ANALELE ŞTIINŢIFICE ALE UNIVERSITĂŢII „ALEXANDRU IOAN CUZA" DIN IAŞI
TOM LXX, 2, ŞTIINŢE JURIDICE, 2024

DOI: 10.47743/jss-2024-70-2-3

AI Ethics: The Bias Puzzle

Alexandru CHISTRUGA[1]

Abstract: The advantages of artificial intelligence are extensively discussed in specialized literature, which claim that technology has the power to fundamentally change society. However, rapid development of artificial intelligence does carry some serious risks, the most important of which is the spread of false and discriminatory information. Since artificial intelligence is "fed" with data from many sources, there is an increased risk that some of the data contains extremist or xenophobic literature. In such circumstances, artificial intelligence could spread extremely dangerous theories and ideas. Thus, government intervention is required to preserve control over the different data categories that developers have access to. As an example, we would like to bring up the fact that during testing, one of the most well-known AI interfaces, GPT-4, provided "advice" on how to murder a huge amount of people for a single dollar and what messages to promote in order to attract people to join Al-Qaeda.

Keywords: artificial intelligence; disinformation, data.

Introduction

Artificial intelligence is constantly evolving, making it difficult to find a field in which it is not used. For instance, in medicine, AI is used *„to help process medical data and give medical professionals important insights, improving health outcomes and patient experiences"*[2]. In some cases, the analysis provided helps clinicians to detect the onset of a disease in a timely manner, including situations in which artificial intelligence has proven to be more efficient than professionals in the area. As a rule, *„the most common role for AI in medical settings are clinical decision support and imaging analysis*[3]*"*. The initial phase involves examining the data concerning a specific patient's case, followed by correlating the obtained information with prior knowledge, ultimately providing a response that may manifest in the form of a treatment proposal.

[1] PhD Student, Faculty of Law, „Alexandru Ioan Cuza" University of Iaşi, e-mail: alexandruchistruga98@gmail.com

[2] IBM, *What is artificial intelligence in medicine?*, online at https://www.ibm.com/topics/artificial-intelligence-medicine#:~:text=Artificial%20intelligence%20in%20medicine%20is%20the%20use%20of,important%20insights%2C%20improving%20health%20outcomes%20and%20patient%20experiences, accessed on 08.05.2024.

[3] *Ibidem.*

Until recently, it was considered that artificial intelligence would eventually remain a robot incapable of creativity, preventing a part of the occupations from being replaced. But, *„a recent study suggests that large language model artificial intelligence chatbots excel beyond the average human in creative tasks"*[4]. Moreover, another study shows that business is anticipating *„that new technologies will destroy jobs faster that creating new ones over the next five years – with a net negative of 14 million roles".* For these reasons, some of the most important players in the IT sector, among them Elon Musk and Steve Wozniak, urged *„all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4"*[5]. Some of these concerns are legitimate, but in this work, we will focus on a lesser-known risk: the spread of discriminating or biased responses by artificial intelligence.

1. Exploring AI Biases

Most artificial intelligence systems, such as ChatGPT, provide, in addition to useful information, a set of responses that reflect and perpetuate human biases, which can lead to incorrect findings in fields such as medicine, human resources, and justice[6]. In specialized literature, biases are defined as an *„effect that deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error, which may distort on any one occasion but balances out on the average"*[7]. Typically, this category includes discriminatory practices that have resulted in unequal treatment of individuals in identical situations based on traits like gender or race.

The reasons behind the emergence of biases have a direct connection to the data that artificial intelligence-based platforms are able to obtain[8]. In this

[4] *New Study: AI Chatbots Surpass the Average Human in Creativity,* SciTechDaily, 15 September 2023, online at https://scitechdaily.com/new-study-ai-chatbots-surpass-the-average-human-in-creativity/#:~:text=A%20recent%20study%20published%20in%20the%20journal%20Scientific,common%20items%20%E2%80%93%20a%20reflection%20of%20divergent%20thinking, accessed on 08.05.2024.

[5] Future of life Institute, Pause Giant AI Experiment: An Open Letter, 22 Marth 2023, online at https://futureoflife.org/open-letter/pause-giant-ai-experiments/, accessed on 08.05.2024.

[6] *Declaration on Ethics and Data Protection In Artificial Intelligence,* 40 th International Conference of Data Protection and Privacy Commissioners, 23 October 2018, online at: https://www.privacyconference2018.org/system/files/2018-10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf, accessed on 08.05.2024.

[7] R. Schwartz, A. Vassilev, L. Perine, A. Burt, P. Hall, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence,* NIST Special Publication 1270, online at https://doi.org/10.6028/NIST.SP.1270, accessed on 08.05.2024.

[8] M. Rijmenam, *Privacy in the Age of AI: Risks, Challenges and Solutions,* The Digital Speaker, 17 February 2023, online at https://www.thedigitalspeaker.com/privacy-age-ai-risks-challenges-solutions/, accessed on 12.05.2024.

regard, for its development, AI is trained on a large amount of data that is specific to the field in which it is used[9]. As a rule, artificial intelligence system developers should collect only representative data. However, in practice, they *„amass their training sets through automated tools that catalog and extract data from the Internet"*[10]. Simply put, AI has access to any kind of content, including pirated books[11], public Facebook and Instagram posts[12] or information from Wikipedia or Reddit. Despite the fact that the vast majority of the information provided is correct, there is a risk of „infiltrating" a set of fake information[13]. So, in some circumstances, artificial intelligence is „fed" data that does not match reality, perpetuating false responses[14].

AI biases are classified into a variety of categories, the most prevalent of which are systemic. These types of biases occur unintentionally as a result of institutional policies that favour particular social categories. At the same time, biases may arise as a result of incorrect analysis of data sources used to train artificial intelligence systems. One example in this sense are the results provided by image-generating platforms, which associate men with the most well-paid jobs, such as doctors or programmers, while women are associated with activities particular to previous centuries, such as housekeepers. In this regard, Stable Diffusion, a platform that converts text into images, has generated images in which women are underrepresented in the majority of industries. Thus, *„women made up a tiny fraction of the images generated for the keyword judge – about 3% – when in reality 34% of US judges are women, according to the National Association of Women*

[9] ET Online, *AI and Privacy: The privacy concerns surrounding AI, its potential impact on personal data*, The Economic Times, 25 April 2023, online at https://economictimes.indiatimes.com/news/how-to/ai-and-privacy-the-privacy-concerns-surrounding-ai-its-potential-impact-on-personal-data/articleshow/99738234.cms?from=mdr, accessed on 12.05.2024.

[10] L. Leffer, *Your Personal Information Is Probably Being Used to Train Generative AI Models,* Scientific American, 19 October 2023, online at https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/, accessed on 12.05.2024.

[11] A. Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative AI*, The Atlantic, 19 august 2023, online at https://www.theatlantic.com/ technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/, accessed on 12.05.2024.

[12] K. Paul, *Meta s new AI assistant trained on public Facebook and Instagram posts*, Reuters, 29 september 2023, online at https://www.reuters.com/technology/metas-new-ai-chatbot-trained-public-facebook-instagram-posts-2023-09-28/, accessed on 13.05.2024.

[13] M. Khatri, *Data Privacy in the Age of Artificial Intelligence (AI),* Linkedin, 18 august 2023, online at https://www.linkedin.com/pulse/data-privacy-age-artificial-intelligence-ai-mousam-khatri, accessed on 13.05.2024.

[14] R. Healey, *Data Privacy Compliance a Significant Challenge to AI Technology*, Formiti Data International, online at https://formiti.com/data-privacy-compliance-a-significant-challenge-to-ai-technology/, accessed on 16.05.2024.

Judges and the Federal Judicial Centre"[15]. This example demonstrates that historical data takes precedence over current statistics, with artificial intelligence prioritizing quantity over quality of data. In other words, artificial intelligence is now incapable of distinguishing between historical data and reality.

Apart from gender-related biases, artificial intelligence also generates erroneous replies based on the race of the individuals engaged[16]. Hence, Stable Diffusion „*generated images of people with darker skin tones 70% of the time for keyword fast-food worker, even though 70% of fast-food workers in the US are white*"[17]. At the same time, „*more than 80% of the images generated for the keyword inmate were of people with darker skin, even though people of colour make up less than half of the US prison population, according to the Federal Bureau of Prisons*"[18]. Therefore, once again, biases occur because AI does not take into account the statistical data currently available, which leads to the appearance of biases[19].

This situation is not unique to the Stable Diffusion platform. Even now, Midjourney and Dall-e, the two most popular text-to-image AI platforms, still generate biased images. In a test, the researcher requested Midjourney to generate images of Barbie dolls, each of which had to represent a specific state. The outcomes were full of biases, „*several of the Asian Barbies were light-skinned, Thailand Barbie, Singapore Barbie, and the Philippines Barbie all had blonde hair, and Germany Barbie wore military-style clothing*"[20]. Similar results were obtained when the platform was asked to submit 100 images of citizens living in particular states. In this regard, „*an Indian person is almost always an old man with a beard, a Mexican person is usually a man in a sombrero, while American person appeared to be overwhelmingly portrayed by the presence of U.S. flags"*[21].

While the number of biases perpetuated by artificial intelligence is immense, we will focus on those that already affect us directly, notably in domains such as recruitment, public security, and medicine. Furthermore, we will discuss results from an experiment that demonstrated humans' predisposition to embrace and spread AI-generated incorrect responses.

---

[15] L. Nicoletti, D. Bass, *Humans are Biased. Generative AI is even worse,* Bloomberg Technology, 9 June 2023, online at https://www.bloomberg.com/graphics/2023-generative-ai-bias/, accessed on 17.05.2024.

[16] T.J. Thomson, R.J. Thomas, *Ageism, sexism, classism and more: 7 examples of bias in AI-generated images*, The Conversation, 10 iuly 2023, online at https://theconversation.com/ageism-sexism-classism-and-more-7-examples-of-bias-in-ai-generated-images-208748, accessed on 17.05.2024.

[17] L. Nicoletti, *op. cit.*

[18] *Ibidem.*

[19] S. Kapoor, A. Narayanan, Quantifying ChatGPT's gender bias, AI Snake, 26 April 2023, online at https://www.aisnakeoil.com/p/quantifying-chatgpts-gender-bias, accessed on 18.05.2024.

[20] V. Turk, *How AI reduces the world to stereotypes,* Rest of World, 18 October 2023, online at https://restofworld.org/2023/ai-image-stereotypes/, accessed on 19.05.2024.

[21] *Ibidem.*

1.1. AI Biases in Recruitment: Challenges and Implication

As noted earlier, artificial intelligence is fed a vast amount of data and is capable of processing all the information really quickly. In light of these capabilities, a number of companies have begun developing systems that let them streamline the hiring process.

As a rule, the recruitment process involves several stages, including the search phase, screening, interviews, and candidate selection. The search phase can be almost entirely automated, with artificial intelligence capable of analyzing the profiles of candidates who have posted their resumes on recruitment websites. Following the analysis of public data, artificial intelligence could suggest to the company to make job offers to candidates who best match the established requirements. At the same time, artificial intelligence could also be used in the screening phase, where it would analyse candidate profiles to identify the most suitable person for a particular position. In both cases, at least in theory, the human factor that might reject candidates due to biases against a certain group of people would be eliminated. Furthermore, since artificial intelligence can review hundreds of resumes in a short amount of time, it would take less time to review candidate profiles.

Artificial intelligence may suggest hiring some applicants over others without adequate justification. For example, Amazon developed a platform to review candidate resumes for specific technology industry positions[22]. The computer models „*were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period*," most of which came from men[23]. As a result, the artificial intelligence „learned" that male candidates were preferable, rejecting resumes submitted by women. This behaviour can be explained by the quality of the data Amazon used to develop the platform, which was unrepresentative. In other words, because the platform was trained with information that no longer reflects reality, it adopted biased behaviour, even though none of the parties involved intended for this to happen[24].

1.2. AI Biases in Healthcare and Public Security

Public security is another area where artificial intelligence may be used. For example, officials in the United States identify people who might be involved in criminal behaviour using facial recognition technology. This technology is „*an artificial intelligence-powered technology that tries to confirm the identify of a person*

[22] IBM Data and AI Team, *Shedding light on AI bias with real world examples*, 16 October 2023, online at Shedding light on AI bias with real world examples – IBM Blog, accessed on 19.05.2024.

[23] J. Dastin, *Insight – Amazon scraps secret AI recruiting tool that showed bias against women*, Reuters, 11 October 2018, online at https://www.reuters.com/article/idUSKCN1MK0AG/, accessed on 20.05.2024.

[24] C. Kerry, *Protecting privacy in an AI-driven world,* Brookings, 10 February 2020, online at https://www.brookings.edu/articles/protecting-privacy-in-an-ai-driven-world/, accessed on 20.05.2024.

from an image"[25]. So, the mechanism involves providing a description of the potential offender to the artificial intelligence system, which uses databases maintained by authorities to generate a list of possible suspects. Due to the overrepresentation of people of colour in law enforcement databases, artificial intelligence sometimes provides incorrect answers, identifying these individuals as potential suspects even when the police officers do not specify the race of the offenders. Consequently, there are situations where individuals with no connection to the investigated case are detained. Because of these kinds of errors, cities like Boston and San Francisco have banned the use of facial recognition technology to identify criminals[26].

If in the case of databases used for training facial recognition technologies, the category of people of colour is overrepresented, in the case of platforms used in the medical field, the situation is diametrically opposite[27]. Thus, in the United States, the databases provided for the development of platforms used in hospitals come from three states, namely California, Massachusetts, and New York, and these are not representative, with only 8.7% of respondents reporting their race and ethnicity. The same situation is valid in the United Kingdom, where out of 500,000 patients, only 6% are non-European. As a result, the use of artificial intelligence is not effective when called upon to provide answers regarding patients from categories underrepresented in the databases with which it was trained.

For example, artificial intelligence is used to identify melanoma, a form of skin cancer, and it performs quite well when the images provided are of white individuals. However, the chances of melanoma being identified in Hispanics and people of colour are reduced, which can be explained by the lack of databases in which these categories of people are represented in a sufficient numbers[28].

In this regard, the specialized literature refers to the „Asan" database and the "Dermofit" database to demonstrate that the underrepresentation of certain categories of people can lead to a decrease in the accuracy with which artificial intelligence identifies melanoma[29]. Thus, the „Asan" database contains images

[25] T.L. Johnson, N.N. Johnson, *Police Facial Recognition Technology Can't Tell Black People Apart,* Scientific American, 18 May 2023, online at: https://www.scientificamerican.com/article/police-facial-recognition-technology-cant-tell-black-people-apart/, accessed on 20.05.2024.

[26] *Ibidem.*

[27] T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, et. al., *Accessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study*, The Lancet Digital Health, Volume 6, Issue 1, E12-E22, January 2024, online at: https://doi.org/10.1016/S2589-7500(23)00225-X, accessed on 21.05.2024.

[28] J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparties in Commercial Gender Classification,* Proccedings of Machine Learning Research, 81: 1-15, Conference on Fairness, Accountability, and Transparency, January 2018, online at https://proceedings.mlr.press/v81/buolamwini18a.html, accessed on 21.05.2024

[29] M. Gayal, *Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities*, Computers in Biology and Medicine, Volume 127,

specific to the Asian population, while the "Dermofit" database contains images specific to Caucasians. The authors of the study asked the artificial intelligence to identify people from Asia who are predisposed to melanoma, with a result of 80% for platforms that used the „Asan" database and only 56% for the platform that was trained on the „Dermofit" database[30]. Through this study, it was demonstrated that artificial intelligence cannot apply the algorithms it uses to identify medical conditions in one group of patients to other categories of people for which it does not have sufficient data. However, artificial intelligence does not refuse to provide an answer, citing a lack of data.

1.3. Perpetuating AI-Generated Biased Responses

Unfortunately, even when artificial intelligence makes recommendations that are obviously incorrect, people still have a tendency to follow them. In an experiment[31], 169 students had to decide if the people who appeared in the images provided had Lyndsay syndrome or not. The students were split up into two groups. The first group was assisted by artificial intelligence, while the other group had to make decisions without external help. In order to find out if student decisions may be influenced, artificial intelligence was programmed to provide 10 incorrect answers[32]. As anticipated, the artificial intelligence-assisted students made more errors and provided incorrect answers far more frequently than the students in the other control group.

The second experiment involved dividing the students into two distinct groups again, but the novelty was adding 25 additional images for the students to analyse independently. Thus, in the first phase, the conditions were similar to those of Experiment 1, with one group of students assisted by artificial intelligence and another not, but in the second phase both groups were placed in equal conditions, with the participation of artificial intelligence excluded. The purpose of this experiment was to see if students would repeat the incorrect answers that artificial intelligence had recommended in similar contexts. As before, the students who received artificial intelligence assistance produced lower-quality findings than the other group, indicating that people are more likely to spread mistakes that are made while interacting with AI.

December 2020, online at: https://doi.org/10.1016/j.compbiomed.2020.104065, accessed on 22.05.2024

[30] *Ibidem.*

[31] L. Vicente, H. Mature, *Humans inherit artificial intelligence biases*, Scientific reports, 3 october 2023, online at https://www.nature.com/articles/s41598-023-42384-8, accessed on 22.05.2024.

[32] Duesto University, *Trapped in a Dangerous Loop: Humans Inherit Artificial Intelligence Biases*, SciTechDaily, 3 October 2023, online at https://scitechdaily.com/trapped-in-a-dangerous-loop-humans-inherit-artificial-intelligence-biases/#:~:text=People%20can%20adopt%20biases%20from%20artificial%20intelligence%20in,%28systematic%20errors%20in%20AI%20outputs%29%20in%20their%20decisions, accessed on 23.05.2024.

In the last experiment, both groups used artificial intelligence to answer 80 questions divided into two sets of 40 questions each. In the first phase, artificial intelligence supported one group, while in the second phase, it was utilized by the other group. In other words, the researchers wanted to verify if students who were not initially assisted by artificial intelligence would provide more correct answers compared to those who were assisted from the beginning. The results demonstrate, once again, that humans are predisposed to perpetuate the erroneous responses suggested by artificial intelligence[33]. Thus, students who were assisted by artificial intelligence in the first phase continued to make errors in the second phase, unlike the other group of students who made errors only when assisted by artificial intelligence.

Summarizing the points presented, we appreciate that we have managed to demonstrate that artificial intelligence can unintentionally provide responses containing erroneous information. Furthermore, the last example also shows how humans are inclined to have too much trust in artificial intelligence, adopting its logical errors. This fact is extremely concerning, especially because artificial intelligence is widely used in fields such as medicine and law enforcement. Consequently, there are cases in which the answers given by artificial intelligence violate individual freedoms or make it more difficult for them to get the care they need[34]. However, artificial intelligence cannot be entirely held liable since humans are the ones who make the final decisions.

At the same time, we chose to present as many examples from different domains as possible to demonstrate that the premises leading to biased responses are diverse, making it extremely difficult to identify a universal solution. However, a common element can be identified, namely the existence of a disproportionality in the information from the databases used to train artificial intelligence. In this regard, the suggestion of lower-paying professions for women and people of colour is the result of the prevalence of historical data suggesting that white men have held professions such as judges or directors. On the other hand, providing erroneous responses regarding the identification of diseases such as skin cancer is the result of the underrepresentation of certain categories of people. In other words, the lack of balance in the information with which artificial intelligence is trained leads to the perpetuation of biases and the provision of wrong solutions.

[33] L. Leffer, *Humans Absorb Bias from AI—And Keep It after They Stop Using the Algorithm*, Scientific American, 26 October 2023, online at https://www.scientificamerican.com/article/humans-absorb-bias-from-ai-and-keep-it-after-they-stop-using-the-algorithm/, accessed on 23.05.2024.

[34] H. Zhang, A.X. Lu, M. Abdalla, M. McDemott, M. Ghassemi, *Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings*, ACM Conference on Health, Inference and Learning, 11 Marth 2020, online at https://doi.org/10.1145/3368555.3384448, accessed on 24.05.2024.

2. Exploring Potential Legislative Solutions for AI Biases

In the first section of this paper, we addressed a number of concerns that artificial intelligence could bring to society, with a particular focus on the biased answers it may generate across a variety of fields. These potential concerns have already caught the attention of national authorities in a number of states, leading to the adoption of artificial intelligence-related regulation. We opted to investigate legislative measures developed in two of the world's main economies, the European Union and the United States of America, to evaluate potential actions taken, with a particular focus on whether regulations address or not the issue of biased responses.

2.1. EU Strategies for Mitigating AI Biases

Most of the risks mentioned in the previous section have been acknowledged by the bodies of the European Union[35], leading to the development of the of the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts[36]. For example, according to recital 18, „*the use of AI systems for 'real-time' remote biometric identification of natural persons in publicly accessible spaces for the purpose of law enforcement is considered particularly intrusive in the rights and freedoms of the concerned persons, to the extent that it may affect the private life of a large part of the population, evoke a feeling of constant surveillance, and indirectly dissuade the exercise of the freedom of assembly and other fundamental rights*". So, the European Union seeks to restrict the use of artificial intelligence in sensitive domains like public space surveillance.

However, according to Article 5, recital 1, letter d), „*the use of real-time remote biometric identification systems in publicly accessible spaces*" is permitted for targeting searches „*for specific potential victims of crime, including missing children*" or for „*the prevention of a specific, substantial, and imminent threat to the life or physical safety of natural persons or of a terrorist attack.*" The use of artificial intelligence to locate victims of crimes or prevent threats could be beneficial, especially considering that authorities currently use various methods involving surveillance of publicly accessible spaces, such as consulting recordings from CCTV cameras. Furthermore, unlike the United States, artificial intelligence is intended to be used for victim identification, with situations where erroneous

[35] European Commision, *AI Act*, Shaping Europe's digital future, online at https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, accessed on 25.05.2024.

[36] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, online at: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF, accessed on 26.05.2024.

answers based on biased information can be provided being limited. In this regard, most likely, artificial intelligence will obtain data characterizing the victim, such as a photo or a sufficiently detailed description, with the platform focusing only on finding a clearly identified person.

The situation is diametrically opposite in the second situation, governed by Article 5(1)(d) of the proposed regulation. The legal provision addresses two unique situations: the use of artificial intelligence to prevent specific, significant, and imminent dangers to people's lives or physical safety, and the use of artificial intelligence to prevent a terrorist attack. In the last case, artificial intelligence will most likely be deployed in scenarios where there is sufficient proof that a terrorist act is likely to occur, with artificial intelligence serving just as a tool to streamline the work of the participating authorities.

We appreciate that it would be extremely difficult for artificial intelligence to be used for such a purpose, as the data provided would in most cases be extremely vague, not allowing for the clear identification of individuals who may be preparing a terrorist attack, and there is also the risk that artificial intelligence may provide biased suggestions. In this regard, the training data provided to artificial intelligence could indicate that individuals with a certain appearance have most often committed terrorist attacks[37]. As a result, there is a possibility that individuals identified by artificial intelligence may have no connection to potential terrorist acts, resulting in the misallocation of human resources. We believe that artificial intelligence should only be used once potential terrorists have been identified through other methods, with the platform being used to ensure their traceability.

Regarding the second case covered by Article 5(d)(2) of the Proposed Regulation, we recognize that the legal provision's formulation is exceedingly ambiguous, providing a danger of abuse by authorities. Thus, under the concept of specific dangers to people's lives or physical safety, any violent acts or activities may be included, and it is unclear how artificial intelligence may be utilized to prevent them. In this sense, the real-time remote biometric identification system has been defined as a *„system whereby the capturing of biometric data, the comparison, and the identification all occur without a significant delay. This comprises not only instant identification but also limited short delays in order to avoid circumvention"*. We believe that artificial intelligence will have access to surveillance cameras, allowing it to monitor in real time if there is cause for a specific action that could be categorized as a threat to individuals. Such 'surveillance' should be authorized in advance by a judicial or administrative authority.

It is difficult for us to identify a situation where compelling reasons could be presented to allow artificial intelligence access to devices collecting biometric data for the purpose of preventing acts of violence. Perhaps the provision provided

[37] A. Abid, M. Farooqi, J. Zou, *Persistent Anti-Muslim Bias in Large Language Models,* AAAI/ACM Conference on AI, Ethics, and Society, 14 January 2021, online la https://doi.org/10.1145/3461702.3462624, accessed on 26.05.2024.

in Article 5 (d) (2) of the Proposed Regulation could be used in a situation where competent authorities have information regarding the preparation of a potential assassination and artificial intelligence would be used to timely identify suspects. However, similar to the prevention of terrorist attacks, the use of artificial intelligence could only be useful if there is sufficient data regarding the individuals who need to be identified; otherwise, there is a risk of providing erroneous responses.

In addition to the above, the Proposed Regulation mentions that artificial intelligence systems which are „*used in education or vocational training, notably for determining access or assigning persons to educational and vocational training institutions or to evaluate persons on tests as part of or as a precondition for their education should be considered high-risk, since they may determine the educational and professional course of a person's life and therefore affect their ability to secure their livelihood*". Even though a series of highly dangerous risks have been identified, the use of artificial intelligence in the field of education will not be prohibited. Classifying these systems as having a high level of risk is intended to limit potential disadvantages.

Developers of artificial intelligence systems will be required to comply with the requirements set out in Article 9 of the Proposed Regulation, including periodic testing of systems to identify, eliminate, or reduce risks. In close connection with the theme of this article, it is explicitly stated that „*training, validation, and testing data sets shall be relevant, representative, free of errors, and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on whom the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.*" We consider this legal provision beneficial, as it eliminates the risk of biased responses.

Meanwhile, the legal provision is extremely restrictive, which could slow down the development of artificial intelligence in the European Union. In other words, developers are required to identify representative databases for each Member State separately, which is why we believe that AI systems will only be developed in those states where compliance with the rules set out in the proposed regulation would be financially justified.

The same conclusions apply to AI systems used in „*employment, workers management and access to self-employment, notably for the recruitment and selection of persons, for making decisions on promotion and termination and for task allocation, monitoring or evaluation of persons in work-related contractual relationships, should also be classified as high-risk, since those systems may appreciably impact future career prospects and livelihoods of these persons".* We have already demonstrated that attempts to use artificial intelligence for personnel recruitment have not been successful, as the solutions offered have been influenced by unrepresentative databases[38].

[38] N. Hanacek, *There's More to AI Bias Than Biased Data, NIST Report Highlights*,

Furthermore, in the field of workforce management, the notion of representative databases should be nuanced. In this regard, there are sufficient professions where the number of male employees outweighs the number of female employees, and vice versa. Training artificial intelligence on databases where the number of employees of a certain gender prevails could lead to the promotion of biases, with Amazon's example being illustrative. For these reasons, we believe that artificial intelligence may not be effectively used in the field of employee recruitment. One possible solution could be to program artificial intelligence to identify one candidate from different categories based on a set of predefined criteria, such as race or gender, allowing the employer to decide which of the suggested individuals best meets the requirements.

2.2. U.S. Initiatives to Combat AI Biases

Regarding the United States of America, on October 30, 2023, President Biden signed the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[39]. In its preamble, it explicitly states that the irresponsible use of artificial intelligence could „*exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security".* At the same time, artificial intelligence systems „*have reproduced and intensified existing inequities, caused new types of harmful discrimination, and exacerbated online and physical harms*".

Therefore, the competent authorities in the U.S. have identified the main risks generated by artificial intelligence, which is why they proposed the adoption of a legislative act aimed at eliminating or at least reducing them. Unlike the regulation adopted at the European Union level, the executive order explicitly addresses the most risky areas where artificial intelligence can have more negative effects than positive ones.

For instance, in the medical sector, as we presented earlier in this paper, artificial intelligence provides a series of biased responses, leading either to the establishment of inappropriate treatment or to limiting the right to access quality services[40]. To mitigate these disadvantages, competent authorities have to develop

National Institute of Standards and Technology, 16 Marth 2022, online at https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights, accessed on 25.05.2024.

[39] WH.GOV, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,* Presidential Actions, 30 October 2023, online at https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence /, accessed on 25.05.2024.

[40] C. Grant, Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism, ACLU Speech, Privacy, and Technology Project, 3 October 2022, online at https://www.aclu.org/news/privacy-technology/algorithms-in-health-care-may-worsen-medical-racism, accessed on 25.05.2024.

„a strategic plan that includes policies and frameworks – possibly including regulatory action, as appropriate – on responsible deployment and use of AI and AI-enabled technologies in the health and human services sector (including research and discovery, drug and device safety, healthcare delivery and financing, and public health)". In this regard, the strategies to be adopted must ensure the use of representative databases and the exclusion or reduction of biased and discriminatory responses, both by current systems and those under development. At the same time, special attention is paid to identifying the erroneous responses provided by artificial intelligence at the present time.

Indeed, the reduction of discrimination and bias is likely to be achieved through the development of representative databases tailored to the intended purpose. We acknowledge that there is a risk that the development of artificial intelligence systems may not proceed as rapidly, as the quantity of training data will be significantly reduced. At the same time, it remains to be seen how developers of artificial intelligence systems will eliminate the unrepresentative data that has already been used to train the platforms. By „extracting" databases, it is likely that some systems will experience a downgrade, requiring them to restart their training processes.

In addition to the medical sector, the executive order stipulates that competent authorities must develop regulatory acts to govern the use of artificial intelligence in areas such as recruitment, access to credit, and other domains where the technology's impact could lead to discrimination.

We appreciate that the legislation developed in the United States will likely respond better to the risks posed by artificial intelligence, as it attempts to directly address the most significant disadvantages, unlike the Regulation adopted at the European Union level, which has established a series of domains in which artificial intelligence cannot be used and has stipulated a series of restrictive obligations that platform developers must adhere to. On the other hand, the regulatory acts mentioned in the executive order have yet to be written, so creators of artificial intelligence-based systems are uncertain of how restrictive the legal restrictions would be, leading to hesitancy in making choices. The Regulation established at the European Union level is scheduled to go into effect in June 2024, giving predictability for those affected, who will be able to adjust considerably more rapidly to the new requirements.

Conclusions

Artificial intelligence has a significant impact on society, with the number of benefits, in our opinion, outweighing the existing disadvantages. However, both in specialized literature and in the public sphere, a series of risks have been presented that indeed seem justified. In this paper, we chose to focus only on the replication and spread by artificial intelligence of biased and discriminatory

responses in various fields, such as medicine or public security[41]. From what has been presented, it emerges that there are situations in which the use of artificial intelligence has had a negative impact on the fundamental rights of the individuals involved, limiting either their freedom or their right to access quality medical services.

The mentioned risks have been recognized by both AI system developers and European and American legislators, leading to the adoption of several regulatory acts directly governing the use of artificial intelligence. The adopted legal provisions are likely to have a negative impact on the development of artificial intelligence. In this regard, imposing restrictions on the quality of data used to train artificial intelligence systems will primarily result in additional costs. Developers will be required to identify representative databases for the sector in which artificial intelligence is to be used, no longer being able to simply access unverified information resources. Consequently, the operating costs of artificial intelligence will increase, likely resulting in its use only in sectors that offer a sufficiently high return on investment to justify the costs. At the same time, cost escalation implies limiting the number of companies that could participate in the development of the sector, potentially leading to its monopolization by Big Tech.

However, we appreciate that adopting regulatory acts in the early stages of artificial intelligence development is beneficial, as it eliminates the risk of AI system developers having to start the process from scratch in the future. In any case, the subjects targeted by the Regulation adopted at the European Union level have had access to relevant information since 2021, having sufficient time to adapt. Due to the new regulations, both existing platforms and those yet to emerge will offer far fewer biased and discriminatory responses, making artificial intelligence an extremely useful tool for society.

References

Abid A., Farooqi M., Zou J., *Persistent Anti-Muslim Bias in Large Language Models,* AAAI/ACM Conference on AI, Ethics, and Society, 14 January 2021, https://doi.org/10.1145/3461702.3462624.

Schwartz R., Vassilev A., Perine L, Burt A., Hall P., *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence,* NIST Special Publication 1270, https://doi.org/10.6028/NIST.SP.1270.

Zack T., Lehman E., Suzgun M., Rodriguez J.A., et. al., *Accessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study*, The Lancet Digital Health, Volume 6, Issue 1, E12-E22, January 2024, https://doi.org/10.1016/S2589-7500(23)00225-X.

Zhang H., Lu A.X., Abdalla M., McDemott M., Ghassemi M., *Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings,* ACM Conference on Health, Inference and Learning, 11 MAarth 2020, https://doi.org/10.1145/ 3368555.3384448.

[41] Information Commissioners Office (ICO), *Guidance on AI and data protecion,* 15 Marth 2023, online at https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ ?template=pdf&patch=17#link1, accessed on 29.05.2024.

Information Commissioner s Office (ICO), *Guidance on AI and data protecion,* 15 Marth 2023, [Online]

Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, [Online]

WH.GOV, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,* Presidential Actions, 30 October 2023, [Online]