MOHAMMAD BITAR, AHMAD KHALIL, S. ANANDHA KRISHNA RAJ, RUPAL MALIK

Legal Assessment of Bias and Discrimination of AI Tools in Higher Education and Research

Abstract

The use of artificial intelligence (AI) tools in higher education has become increasingly important because of the time and effort savings and the speed of information transfer. However, many ethical and legal challenges make their use in this field a complex issue. Problems such as bias and discrimination that arise from AI Tools require the establishment of a legal system capable of controlling their use in an optimal manner. However, the legal regulation of the use of AI Tools in higher education, especially in the fields of research and data analysis, does not reach the required level. Although many countries have begun to use these tools in higher education and scientific research, the legal framework is still not at the required level. This research attempts to explore the legal and ethical challenges of using AI in higher education and scientific research with the aim of focusing on the importance of developing a legal framework capable of promoting the use of AI Tools in the scientific and educational sectors. The paper highlights the most important relevant laws in technologically advanced countries in general to measure the extent to which they are reflected in reality.

KEYWORDS: AI, data training, higher education, bias, discrimination, research

MOHAMMAD BITAR – PhD in law, VIT-AP University,
ORCID – 0000-0003-4686-7411, e-mail: 1mohammad.bitar@gmail.com
AHMAD KHALIL – PhD in law, Vellore Institute of Technology, School of Law (VITSOL), Chennai, India, ORCID – 0009-0007-0615-9812,
e-mail: ahmad.khalil2020@vitstudent.ac.in
S. ANANDHA KRISHNA RAJ – PhD in law, Vellore Institute of Technology, School of Law (VITSOL), Chennai, India, ORCID – 0009-0001-0177-1689,
e-mail: ahmadkhalil5665@gmail.com
RUPAL MALIK – Assistant Professor, Ramaiah College of Law, Bengaluru, India,
ORCID – 0009-0001-2461-8580, e-mail: rupalmalik@msrcl.org

HTTPS://DOI.ORG/10.36128/PRIW.VI56.896

1 | Introduction

The term "Artificial Intelligence" (AI) was coined by McCarthy in 1956.[1] However, this term has evolved over time, but the major boom occurred in the last decade. AI is currently defined as: "computing systems that can engage in human-like processes such as learning, adapting, synthesizing, self-correction and the use of data for complex processing tasks."[2]

Recent years have witnessed a revolution in technological advancement and the significant use of AI. This progress has contributed to various aspects of life, especially in the field of education. AI has provided many distinctive educational solutions and experiences in the field of education.[3] This highlights the urgent need for countries to achieve greater benefit through the greater introduction and application of AI within the scope of education, especially higher education and scientific research.

AI tools have proven their ability to understand language and create texts with high efficiency in a manner similar to human work. AI tools, such as generative language models, have been characterized by using deep learning techniques in addition to a large database.[4] Thanks to these technologies, AI tools are capable of performing analyses, drawing logical conclusions, and generating high-quality responses, in addition to the ability to understand complex contexts and patterns.[5] Large language models have emerged as one of the most important of these tools in the field of higher education and scientific research. This is due to the ability of these tools to create, coherent, and logical text.[6]

[1] Sonia Jawaid Shaikh, "Artificially intelligent, interactive, and assistive machines: A definitional framework for intelligent assistants" *International Journal of Human–Computer Interaction*, No. 4 (2023): 776-789.

[2] Stefan A.D. Popenici, Sharon Kerr, "Exploring the impact of AI on teaching and learning in higher education" *Research and practice in technology enhanced learning*, No. 1 (2017): 2

[3] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, Eamonn Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances" *Data mining and knowledge discovery*, 31 (2017): 606-660.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, "A neural probabilistic language model" *Advances in neural information processing systems*, 13 (2000).

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean, "Distributed representations of words and phrases and their compositionality" *Advances in neural information processing systems*, 26 (2013).

[6] Yang Kai-Cheng, Ferrara Emilio, Menczer Filippo, "Botometer 101: Social bot practicum for computational social scientists" *Journal of Computational Social Science*, No. 2 (2022): 1511-1528.

With the increasing reliance on large language models and other AI tools in the field of scientific research and higher education, many challenges have emerged. The most prominent of these challenges, which has become one of the most disturbing problems, are prejudices and discrimination. These biases have significant negative impacts on users and society.[7]

Biased conclusions and outputs of AI reflect the biases of its developers or society's discriminatory view against a particular group. Practical experiments with AI tools have proven significant biases in their outputs. These biases targeted customs, gender, and cultural background. Biases included sectors as diverse as criminal justice,[8] employment, and medicine[9]. AI algorithms are also characterized by biases in the field of education, such as biases in learning tests,[10] dropout predictions, predicting exam scores, admission to graduate studies, and writing research texts.

The bias of AI tools in some cases leads to discrimination that is prohibited by law in many cases. Discrimination may target a protected characteristic, such as sex, colour, race, or other social characteristics. In other cases, AI leads to new types of discrimination that are not addressed by law. In some cases, AI outputs reflect discrimination unrelated to legally protected characteristics, leading to unfair outcomes because it involves promoting social inequality, for example, such as recruitment advertisements. In light of these legal gaps, the need arises to create a legal mechanism more capable of combating discrimination resulting from AI practices, both in the education sector and in other critical areas such as criminal justice, employment, medicine, image analysis, and machine translation.

To effectively counteract discrimination and bias facilitated by AI tools, a crucial step involves aligning regulatory frameworks with fundamental legal principles related to discrimination rules and data protection. Upholding human rights amidst the ongoing technological revolution necessitates vigilance in ensuring that monitoring committees are established.

[7] Emilio Ferrara, "GenAI against humanity: Nefarious applications of generative AI and large language models" *Journal of Computational Social Science*, (2024): 1-21.

[8] Duncan N. Angwin, Kamel Mellahi, Emanuel Gomes, Emmanuel Peter, "How communication approaches impact mergers and acquisitions outcomes" *The International Journal of Human Resource Management*, No. 20 (2016): 2370-2397.

[9] Vikas O'Reilly-Shah, Katherine R. Gentry, Wil Van Cleve, Samir M. Kendale, Craig S. Jabaley, Dustin R. Long, "The COVID-19 pandemic highlights shortcomings in US health care informatics infrastructure: a call to action" *Anesthesia & Analgesia*, No. 2 (2020): 340-344.

[10] Rosamond Mitchell, Florence Myles, Emma Marsden, *Second language learning theories* (London: Routledge, 2019).

These committees play a pivotal role in preserving the authenticity and impartiality of data, thereby mitigating potential biases and discrimination introduced by AI tools.

The challenge at hand extends beyond addressing established concepts of discrimination; it also entails the potential creation of novel discriminatory notions by AI, posing a substantial obstacle. Given the inherent complexity of AI, a broad organizational approach may prove insufficient. Therefore, a more focused and purpose-driven classification is imperative. This research aims to delve into the ramifications of bias and discrimination introduced by AI tools in higher education and scientific research. Thorough investigation and analysis are essential to yield insights that can inform a nuanced legal understanding. This understanding, in turn, is crucial for steering regulatory efforts in a direction that ensures robust and effective oversight.

2 Explaining Bias in Generative Language Models

2.1. Key Contributors to Bias in Large Language Models

Recently, language models have contributed significantly to the process of writing research, analysing statistics and data, and generating texts used in writing research and educational texts in higher education. Which led to a high rate of reliance on these tools by researchers and universities. Despite the many positives presented by the language models,[11] their use was a double-edged sword. Some of the texts generated, data analysis, and research results were biased. This bias manifested itself in the form of incorrect attribution of texts, distortion of facts, and bias towards one group or ideas at the expense of others in a way that amounts to discrimination, in addition to emphasizing some stereotypes that do not reflect the truth. These biases arise for a variety of reasons, including algorithms, political decisions, product design decisions, training data, test scoring, and explanatory data. Certain algorithms may biasly inflate some specific points. Political decisions may also play into biasing language models,

[11] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee et al., "Sparks of artificial general intelligence: Early experiments with gpt-4" *arXiv preprint arXiv:2303.12712* (2023).

as developers may resort to supporting or preventing certain results.[12] In addition, product design decisions play in biasing language models by giving priority to usage models. Certain types of content are designed for specific populations and users in a way that leads to unintended bias towards other groups.[13] Training data also plays a significant role in potentially biasing, and when specific training models are chosen in a biased manner, this necessarily leads to lead to biased results.[14] Finally, subjective judgments and individual comments may lead to biased results and misunderstanding of the data by linguistic models due to the linguistic models being affected by these comments.[15]

2.2. Types of biases in Large Language Models Large Language Models

In light of the increasing reliance on language models to assist in writing research texts, analysing data, and providing information in higher education, it is necessary to take into account the biases that result from the outputs of language models.

Language modelling biases in generated texts and data take multiple forms, such as ideological, political, demographic, temporal, and cultural linguistic biases, as well as confirmation biases. Linguistic biases are particularly pronounced for less commonly used languages or minority dialects. For example, the languages of education in most Asian countries are less supported languages compared to the more common languages such as English, which makes these languages less efficient when used in the field of research.[16] Political and ideological biases also appear in the

[12] Finale Doshi-Velez, Been Kim, "Towards a rigorous science of interpretable machine learning" *arXiv preprint arXiv:1702.08608*, (2017).

[13] Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan, "Inherent trade-offs in the fair determination of risk scores" *arXiv preprint arXiv:1609.05807* (2016).

[14] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, Adam T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings" *Advances in neural information processing systems*, 29 (2016).

[15] Joy Buolamwini, Gebru Timnit, "Gender shades: Intersectional accuracy disparities in commercial gender classification" *Proceedings of the 1st Conference on fairness, accountability and transparency*, 81 (2018): 77-91.

[16] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou, "Word translation without parallel data" *arXiv preprint arXiv:1710.04087* (2017).

results of linguistic models, reflecting the biases present in the training data for these models. The generated texts may deepen existing political and ideological biases.[17] Existing demographic biases appear through outputs that show biased behaviour toward norms, races, minorities, or social groups because they are based on training data that does not reflect the true size of these demographic groups.[18] Temporal biases arise because the training data on which language models are based is limited to a certain period of time and therefore the outputs are usually biased with respect to more recent opinions and trends. On the other hand, language models are usually less able to understand historical contexts due to the lack of training data.[19] Language model biases also take the form of cultural biases, as in many cases the output of language models may perpetuate stereotypes and pre-existing cultural biases due to their reliance on culturally biased training data.[20] The last type of biases that may result from the output of language models. These are confirmation biases that reinforce individuals' beliefs by providing outputs that are consistent with their views. This occurs when individuals use language models to search for information that is consistent with their ideas, and these are unintentional biases.[21]

2.3 The Influence of Training Data Sources on Bias in Large Language Models

Large language models like Chat GPT rely on unsupervised machine learning processes. These models learn based on large amounts of unlabelled data. Websites, books, articles, texts on social networking sites, and other

[17] Dixon Lucas, Li John, Sorensen Jeffrey, Thain Nithum, Vasserman Lucy, "Measuring and mitigating unintended bias in text classification", [in:] *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018), 67-73.

[18] Aylin Caliskan, Joanna Bryson, Narayanan Arvind, "Semantics derived automatically from language corpora contain human-like biases" *Science*, No. 6334 (2017): 183-186.

[19] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi, "Defending against neural fake news" *Advances in neural information processing systems* (2019).

[20] Shikha Bordia, Samuel R. Bowman, "Identifying and reducing gender bias in word-level language models" *arXiv preprint arXiv:1904.03035*, (2019).

[21] Caliskan, Bryson, Narayanan, "Semantics derived automatically from language corpora contain human-like biases" *Science*, No. 6334 (2017): 183-186.

written texts available on the Internet are the source of data that large language models such as Chat GPT rely on in training processes.[22]

Sources of training data on which large language models are based include Internet sites including text in news sites, blogs, and information sites such as Wikipedia. In addition to the books on which large language models are based on writing methods, textual narration and various information. Social media platforms are also considered an important source of training data regarding contemporary discussions and colloquial and colloquial languages.

Whereas, although large language models filter the training data before entering it into the model in order to remove random and low-value data,[23] this does not prevent a lot of low-quality content from leaking due to the huge volume of texts [24], Which in turn may lead to biased results.

3 | Revealing the Extent of the Risk of Discrimination in AI

3.1 The Nature of Discrimination Produced by the Use of AI

To detect the risk of discrimination resulting from the use of AI, it is necessary to determine the nature of this problem, which is what we will discuss in this section. AI may result in discrimination due to its inherently ambiguous nature. For this reason, AI is often described as a “black box”.[25]

In scientific research, the researcher may see AI for a specific group of people. The lack of clarity in decision-making often leads to confusion. There is a weakness in the ability to estimate the reason behind the

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “Pre-training of deep bidirectional transformers for language understanding” *arXiv preprint arXiv:1810.04805* (2018): 8.

[23] Emilio Ferrara, “The history of digital spam” *Communications of the ACM*, No. 8 (2019): 82-91.

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, “Language models are unsupervised multitask learners” *OpenAI blog*, No. 8 (2019): 9.

[25] Frank Pasquale, *The black box society: The secret algorithms that control money and information* (Cambridge: Harvard University Press, 2015).

decision-making process as to whether discrimination was for instance made on an ethnic basis or not.

What makes the task more difficult is that AI does not follow a single path in making decisions that may lead to discrimination. Some scholars have concluded that AI can engage in five unintended decision-making approaches that are likely to result in discrimination. To fully understand this issue, it is essential to examine the specific problems that arise in AI decision-making. [26] These problems include: (1) determining the nature of the "variable target" and "class labels," (2) the presence of discriminatory labels in training data, (3) challenges in data collection for AI training, (4) the impact of feature selection, and (5) the role of proxies. Each of these issues contributes to the risk of both unintentional and intentional discrimination in AI systems. The following sections will provide a detailed review of these problems, examining their causes, implications, and potential solutions.

3.1.1. Determining the Variable Target and Class Labels

AI links on computers are based on extensive data. These links sometimes take the form of spam in email when the developers work on the email filtering feature. The process involves tracking how users classify or sort emails on computers so that the developer relies on, which are collected from users, are called training data.

The computer detects spam in emails by determining their characteristics. This process by which characteristics are determined by correlations is called emergent or predictive modelling. This process is embodied, but not limited to, in specifying specific words or phrases that are significant, such as "You have been chosen in the drawing to win a car" or "Publish your article within 10 days in the Scopus Indexed Journal." These specific words or phrases may originate from specific IP addresses. The process of presenting those specific words or sentences by machine learning features produces algorithms that learn about everything that intersects with those activities. These relevant outcomes are called the "target variable".[27]

The target variable is considered a compass for developers in their reliance on the extracted data. Each category is the result of dividing all the

[26] Solon Barocas, Selbst D. Andrew, "Big data's disparate impact" *California Law Review*, 104 (2016): 671.

[27] Ibidem, 678.

values of the extracted data to form interchangeable and exclusive categories.[28] However, people's collective agreement in classifying spam is very important to name this category, so that spam can be filtered on this basis.

In some cases, a kind of ambiguity is created to define the target variable, which calls for expanding the definition of the categories by creating a new type. For example, when developing AI for scientific research. Suppose we study a country or city based on the most criminal neighbourhoods to predict the future rate of taking preventive measures and launching intensive awareness programs. The question that arises here is whether the target variable will be considered a "crime" depending on the guilty mind. Or just based on the percentage of crime and the number of crimes committed? Crime data algorithms can discriminately target specific areas, and this collected data is the result of historical evidence of biased police practices in one or more specific neighbourhoods. The analysis of the target variable will be discriminated and unfair, arising from discriminated historical data. The same applies if a company relies on AI to select workers. The target variable will be for the employee to be "competent." What should the "class designation" of competence be? Does he adhere to work deadlines without any delay? or the one who does not adhere to it but has better productivity?

Therefore, some acknowledge that the intended variables may play a significant or minor negative role in the protected classes, and the same applies to class labels.[29] For example, let us assume that the variable of the target is the criminal act. The environment in which people live varies according to the type of services and care provided, so certain areas may generate a criminal nature or problems for environments with poor services in which ignorance is widespread, and in contrast to the neighbourhoods in which the government cares about health and service awareness which reduces crime from occurring.

Crime has factors resulting from circumstances, not the "criminal mind." The category to evaluate a criminal person depends on the criminal mind without looking beyond the original reason, which is discrimination.[30] Therefore, discrimination may creep into AI as a result of the practices

[28] Ibidem.

[29] Ibidem, 679.

[30] Tabitha C. Peck, Sofia Seinfeld, Salvatore M. Aglioti, Mel Slater, "Putting yourself in the skin of a black avatar reduces implicit racial bias" *Consciousness and Cognition*, No. 3 (2013): 779-787.

of responsible organizations regarding identifying target variables and naming them into categories.

3.1.2. The Discriminatory Labels of the Data in the Field of Training

If discriminatory data is used in training, discrimination will undoubtedly result from AI. According to specialists, there are two methods of training that can result in discrimination by AI. Firstly, through error resulting from problems with the AI system, by choosing a biased sample to train on. Secondly, by training AI naturally by humans on biased data. The result is that there is no doubt that the risk of bias will result from a particular AI system that was trained on biased data or through its error system.

3.1.3. Data Gathering for AI Training

The procedure of collecting biased training samples likely leads to discrimination. If we assume that the AI system is to detect criminals, then if samples are collected to train the AI based on a bias towards a specific race, the AI will result in discrimination. If a particular country practices persecution of a specific region or race, the authorities will document systematic discrimination against them. Moreover, if AI is trained on such collected data, it will lead to discrimination against the oppressed group.[31] For example, ChatGPT, which includes a huge number of words and data. However, these large data sets can be biased by coding and developing social stereotypes to produce discrimination.

For example, in 2022, Chat GPT was trained on data from 48 samples, which led to producing biased information. Those results were cancelled and the biased data exercise was shut down several days later.[32] Another example about Amazon and recruitment, where the company used the AI recruitment program in 2018, but it resulted in discrimination against female candidates.[33]

[31] Lum Kristian, William Isaac, "To predict and serve?" *Significance*, No. 5 (2016): 16.

[32] Jana Sirsendu, Michael R. Heaven, Charles B. Stauft, Tony T. Wang, Matthew C. Williams, Felice D'Agnillo, Abdu I. Alayash, "HIF-1α-Dependent metabolic reprogramming, oxidative stress, and bioenergetic dysfunction in SARS-CoV-2-infected hamsters" *International Journal of Molecular Sciences*, No. 1 (2022): 558.

[33] James D. Hamilton, "Why you should never use the Hodrick-Prescott filter" *Review of Economics and Statistics*, No. 5 (2018): 831-843.

3.1.4. The Effect of the Feature to be Chosen

The importance of choosing the feature or attribute that belongs to the data category is related to simplifying the visualisation in AI. The more limited the feature selection, the higher the risk of bias in AI outcomes. If a company works to choose a specific feature in scientific research for a specific type of information, we will find that AI will be biased toward that feature. Therefore, using predictive features leads to discriminatory effects in AI.

3.1.5. Proxy

This is the case when the training data is protected. Since agents have trust, the criteria for a sound and effective decision are considered an agent. For example, if a proxy creates biased data regarding autonomous weapons[34] stating that these weapon systems are illegal in AI, like Chat GPT. Obtaining the contrary information will not be easily possible due to bias. All information will favour that label category, and all results will show that weapons systems should be banned due to their illegality.

It must be noted finally that discrimination can be done intentionally.[35] for example, when an organization deliberately uses agents to discriminate on an ethnic basis. According to experts, the one in charge of making the decision may skew data in the training field or resort to a specific agent for protected classes, which undoubtedly leads to a discriminatory AI system.[36] The difficulty of detecting discrimination when using a proxy is greater than when using direct discrimination.

[34] Ahmad Khalil, "Development and Deployment of Autonomous Weapon Systems: Comprehensive Analysis of International Humanitarian Law." *Prawo i więź*, No. 3 (2024).

[35] John Bryson, Alessandro Sancino, John Benington, Eva Sørensen, "Towards a multi-actor theory of public value co-creation" *Public Management Review*, No. 5 (2017): 640-654.

[36] David S. Kroll, Harry Reyes Nieva, Arthur J. Barsky, Jeffrey A. Linder, "Benzodiazepines are prescribed more frequently to patients already at risk for benzodiazepine-related adverse events in primary care" *Journal of General Internal Medicine*, 31 (2016): 1027-1034.

3.2. The Use of AI and the Inherent Risks of Discrimination in Certain Areas

We will mention some examples of discrimination resulting from the use of AI or in some areas, the risk of discrimination may result from it.

3.2.1. The Field of Police and Crime Prevention

The most disreputable systems based on AI are "Correctional Offender Management Profiling for Alternative Sanctions" COMPAS. The COMPAS system is used in the US in the field of crime, specifically to provide predictive indicators about the possibility of repeat crimes committed by defendants in the future. The purpose of the COMPAS idea is to assist judges in making precautionary decisions to subject a person to supervision after his release from detention. The surprise is that "COMPAS" is not programmed on a racial basis. Still, some journalists conducted an investigative investigation, the result of which was that COMPAS is biased in favour of whites.[37] The results of an academic discussion about COMPAS indicated that the system's predictions were about 61 percent correct. In addition, what is striking is that the rating of whites was twice lower on the level of risk than blacks, even though blacks, in a large percentage, do not commit future crimes. but in fact, the opposite was true, and the rate of whites' recidivism was greater.[38]

ProPublica confirmed that unfair treatment is very serious. Disproportionality in assigning discriminatory results to specific groups by mistake is unacceptable and puts them at risk. Continuing to monitor a person after his release from prison solely because of the colour of his skin, based on a discriminatory analysis by the AI, is risky. On the contrary, it is more dangerous to give a wrong analysis of the white-skinned person; the reality proves that he is the one who should be monitored because his rate of committing the crime is higher. In summary, discrimination against blacks was not only dangerous in terms of accusing them of crime. But the

[37] Sune Hannibal Holm, Kasper Lippert-Rasmussen, "Discrimination, Fairness, and the Use of Algorithms" *Res Publica*, No. 2 (2023): 177-183.

[38] Angwin, Mellahi, Gomes, Peter, "How communication approaches impact mergers and acquisitions outcomes," 2370-2397.

matter has reached the point of giving false data about who deserves these precautionary decisions.

Justice here must be done at the level of entire groups, not just at one level. If this element is not met, the judge must explain the risk of justice appropriately. The fairness risk in the previous example received different classifications for whites and blacks, indicating the emergence of bias. Some have conducted valuable statistics regarding the occurrence of differences in the tendency to recommit crimes in the future, which makes the mathematically questionable regulation unverifiable concerning the margin of error rate.[39]

3.2.2. In the Field of Analysis and Image Search

Discriminatory effects may result from AI systems that aim to search and analyse images. The development of this model from AI appears on the surface to be very useful, but in reality, it can distort the facts. For example, in Stable Diffusion, a large number of images were created, reaching 5,000, which in the analysis turned out to be discriminatory in nature. This type of AI tool is not limited to Stable Diffusion; for example, platforms such as Dall-E from Open AI and others have a high impact on the future of higher education research.[40] Forgery has also occurred in the production of photos and videos used by opponents of the current US President when they fabricated a film of illegal immigrants and spread widely as being real.[41]

Analyses also showed that the discriminatory classification of images according to gender leads to greater dominance of men in jobs and professions to the exclusion of women, with the exception of some not-so-prestigious professions.

Moreover, men of a certain race, “white,” enjoy higher advantages than others, especially in salaries. In the same context, some keywords that are considered offensive and defamatory are attributed to black people. Finally, there is no doubt that the situation will worsen if the use of generative AI continues unchecked by the criminal justice systems.

[39] Alexandra Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments” *Big Data*, No. 2 (2017): 153-163.

[40] James Brusseau, “AI human impact: toward a model for ethical investing in AI-intensive companies” *Journal of Sustainable Finance & Investment*, No. 2 (2023): 1030-1057.

[41] Timo Schick, Hinrich Schütze, ”It’s not just size that matters: Small language models are also few-shot learners” *arXiv preprint arXiv:2009.07118* (2020).

3.2.3. The Impact of Machine Translation by AI

We are likely to find a distinction when using AI-powered translation tools. For example, some languages, such as Turkish, have a clear male bias that appears when using Google Translate. Confirmation of that, converting the text from Turkish to English brings gender inequality. In the medical field, for example, a woman is translated as a nurse, and a man as a doctor. This is because the language is fundamentally discriminatory. The matter can be explained simply by saying that AI reflects human behaviour and thinking. This is not limited to the Turkish language only. Rather, research was conducted on other languages, and it became clear that the trend reflects a purely masculine approach Prates.[42] In addition, the tendency toward males is exaggerated in many fields, and that has entrenched stereotypes, such as those in natural science, engineering, and mathematics.[43]

In short, AI-based translation tools are likely discriminatory, especially between genders. What was presented as realistic examples can only indicate the danger of using it without meaningful supervision, which may affect human behaviour and thinking.

3.2.4. The Role in Mitigating Risks

Although it has been argued that AI can be discriminatory, in general, AI's behaviour is not expected to be any worse than that of humans.

It cannot be denied that in the real world, humans exhibit discriminatory behaviour. It has also been discussed that AI has discriminatory effects, the most powerful of which is that humans have trained it on discriminatory data. Therefore, it is unfair to compare the distinction between AI and humans in decision-making because humans will undoubtedly prevail.[44] Some have pointed out that AI tools can play a positive role in detecting discrimination. The idea is based on the fact that AI is a reflection of the behaviour and practices of humans. It is possible that if the AI had not revealed

[42] Marcelo Prates, Pedro H. Avelar, Luís C. Lamb, "Assessing gender bias in machine translation: a case study with google translate" *Neural Computing and Applications*, 32 (2020): 6363-6381.

[43] Ibidem.

[44] Tene Omer, Jules Polonetsky, "Taming the Golem: Challenges of ethical algorithmic decision-making" *North Carolina Journal of Law & Technology*, nr 1 (2017): 125.

to us the discrimination, the matter would have remained unknown, as it indirectly sheds light on the negative behaviour of humans.[45]

Nevertheless, this argument does not justify not subjecting AI tools used for scientific research to legal and ethical oversight. Legal regulation of AI tools is necessary to ensure that basic human rights rules are not violated.

4 | Addressing Discrimination in AI: Legal Perspectives and Challenges

The use of AI in the field of research and higher education is not without the risk of bias, which sometimes leads to discrimination in a way that violates fundamental legal principles and human rights standards. This section will address non-discrimination laws with the aim of pointing out the need for laws regulating the use of AI.

Discrimination is generally considered prohibited under a large number of international and European agreements and national legal frameworks, including those of technologically advanced and democratic states. Under Article II of the ICCPR,

> all individuals within its territory and subject to its jurisdiction have the rights recognized in the present Covenant, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

Article 14 of the European Convention on Human Rights also affirms the prohibition of discrimination: "The enjoyment of the rights and freedoms stipulated in this Convention must be guaranteed without discrimination on any basis such as sex, race, colour, language, religion, or political or other opinion. "Or national or social origin, membership of a national minority, property, birth or other status.". It is worth noting that the European Convention on Human Rights prohibits both direct discrimination

[45] Charline Daelman, Katerina Yordanova, "AI through a human rights lens. The role of human rights in fulfilling AI's potential", [in:] *Artificial intelligence and the law* (Cambridge: Intersentia, 2023).

and indirect discrimination. Direct discrimination appears in practices that involve the intention to discriminate between persons on the basis of sex, race, colour, language, religion or other protected characteristics.[46] On the other hand, the European Court of Human Rights and European Union law[47] specify that indirect discrimination is represented by practices that appear neutral at the beginning and do not necessarily entail, in principle, an intention to discriminate against protected characteristics, unless they ultimately lead to discrimination.[48]

Indirect discrimination does not require the presence of criminal intent. Rather, it is sufficient for the results to be discriminatory, which applies to most cases of discrimination in AI that involve discrimination.[49] Most texts generated in large language models contain unintended discrimination against a race, language, religion, or other social characteristic. Therefore, it seems that the chance of unintentional discrimination occurring is much higher than intentional discrimination in the field of AI. Here a legal problem arises, which is the lack of clear rules that address the issue of indirect discrimination resulting from language models. The issue of indirect discrimination seems confusing, especially in the context of proving that the outputs of large language models that appear in principle to be built on a neutral basis have a negative, discriminatory effect against a protected group.

However, the claim that the behaviours or outputs of AI are discriminatory can be refuted by relying on objective justifications according to the European Court of Human Rights"[50] and European Union Law. These justifications are required to be objective and logical. In light of this reality, it seems that determining the practices that It can be considered a matter of discrimination that is not clear and easy.

On the other hand, the protection of non-discrimination laws focuses on known protected categories, such as discrimination against race, colour,

[46] Direct discrimination is defined as follows in Article 2(2)(a) of the Racial Equality Directive 2000/43/EC

[47] ECtHR, Biao v. Denmark (Grand Chamber), No. 38590/10, 24 May 2016, para. 103.

[48] ECtHR, Biao v. Denmark (Grand Chamber), No. 38590/10, 24 May 2016, para. 89.

[49] ECtHR, Biao v. Denmark (Grand Chamber), No. 38590/10, 24 May 2016, para. 103. See also Hacker 2018, p. 1153.

[50] ECtHR, Biao v. Denmark (Grand Chamber), No. 38590/10, 24 May 2016, paras. 91 and 92.

gender, or certain social characteristics.[51] However, there are new types outside these specific categories that may involve discrimination resulting from the behaviour of AI. For example, such as employment advertisements that may not appear to some older people. Many other discriminations resulting from the behaviour of AI are controversial and have not yet been addressed by discrimination laws.

In conclusion, national laws prohibiting discrimination still have many shortcomings to address discrimination resulting from AI practices. In addition to the weakness of enforcement mechanisms, current discrimination laws in the field of AI.

4.1. Evolution of Legal Frameworks

The need to develop laws arises because the current legal framework to prohibit discrimination resulting from AI suffers from many gaps in its application. The approach to developing laws must take two directions: the first focuses on regulating laws in the field of AI in a way that keeps pace with the evolving reality, and the second focuses on the mechanisms for enforcing these laws.

4.1.1. Adapting Legal Frameworks for AI

Creating rules dealing with AI requires updating and creating rules that are flexible and appropriate in a way that keeps pace with technological development and the emerging reality of AI. Where it is possible to start with basic rules and adapt them in a way that is compatible with AI.[52] The basic rules of discrimination are not sufficient to address the evolving technological reality, which requires the creation of special rules that specifically address discrimination resulting from AI practices. Given that the technological reality and AI techniques are constantly evolving, there is a need to conduct a comprehensive evaluation of the rules that regulate the behaviour of AI and modify them whenever necessary.

[51] Tarunabh Khaitan, *A theory of discrimination law* (Oxford: Oxford University Press, 2015).

[52] Bert-Jaap Koops, "Should ICT regulation be technology-neutral?", [in:] *Starting Points for ICT Regulation – Deconstructing Prevalent Policy One-liners*, ed. Bert Jaap Koops et al. (Den Haag: Asser, 2006).

Given that the technological reality and AI techniques are constantly changing and renewed, there is a need to conduct a comprehensive evaluation of the rules that regulate the behaviour of AI and modify them whenever necessary. In order to achieve the establishment of an appropriate legal framework, it is necessary to adopt an approach that combines the established legal rules and guidelines established by regulatory bodies. These guidelines must be easy to modify in a way that enables them to keep pace with subsequent developments in the framework of AI. Regulatory bodies must also follow appropriate legal standards and controls regarding updating flexible guidelines. This hybrid nature of the rules gives greater flexibility in dealing with developments.[53] multi-level legislation with a mixed approach can provide legal stability that regulates AI and subsequent developments.[54]

4.1.2. Improving Transparency and Enforcement in AI Non-Discrimination Standards

In addition to developing a legal framework that regulates the issue of prohibiting discrimination and bias leading to discrimination in the field of AI, there is a need to improve mechanisms for enforcing non-discrimination standards. In light of the existence of general agreement on non-discrimination standards, it seems more confusing in the context of non-discrimination standards in a framework for AI behaviours.[55] The lack of transparency in the decisions and outputs of AI represents the biggest update.[56] Thus, there is a need to improve transparency within the framework of how AI works. It requires that AI algorithms be subject to oversight and interpretation.[57] Transparency is essential to ensuring algorithmic compliance with non-discrimination standards.

In the context of AI systems used in higher education related to testing students or teachers, the algorithm responsible for selection must meet

[53] Ian Brown, Christopher T. Marsden, *Regulating code: Good governance and better regulation in the information age* (Cambridge: MIT Press, 2023).

[54] Koops, "Should ICT regulation be technology-neutral?."

[55] Ahmad Khalil, S. Ananda Krishna Raj, "Challenges to the Principle of Distinction in Cyber Warfare Navigating International Humanitarian Law Compliance: The Principle of Distinction in Cyber Warfare." *Prawo i więź*, No. 2 (2024): 109-131.

[56] Pasquale, *The black box society*.

[57] Aaron Rieke, Miranda Bogen, David G. Robins, *Public scrutiny of automated decisions: Early lessons and emerging methods* (Upturn and Omidyar Network, 2018), 6.

the condition of transparency. For example, it might be a good idea for higher education institutions to share the codes responsible for selecting students and professors in a way that ensures a high level of transparency and compliance with non-discrimination standards. Knowing how AI systems work and the criteria used in making its decisions provides high levels of transparency.[58] This is done by providing the opportunity for concerned persons to evaluate the extent to which these systems comply with non-discrimination standards.

However, transparency requirements often clash with intellectual property rights, privacy protection, and trade secrets.[59] These obstacles make it impossible for concerned persons and academics to assess the extent of compliance with these regulations. This may require the need to ensure that the law requires disclosure of a certain degree of information in a way that enables concerned persons to verify compliance with anti-discrimination standards. This obligation to disclose information must ensure a balance between preserving privacy and trade secrets and transparency, which is necessary to prevent discrimination.

In addition to evaluating the codes and information given, AI systems must be tested on the ground, because examining the codes may not reflect the entire truth about the extent of compliance with non-discrimination standards.[60]

In this context, it becomes clear that while existing legal frameworks highlight important principles, their practical application to AI systems in higher education and research remains fragmented. This creates a pressing need to explore new regulatory models capable of addressing these challenges more systematically.

[58] Mohammad Bitar, Benarji Chakka, "Drone attacks during armed conflict: quest for legality and regulation." *International Journal of Intellectual Property Management* 13, No. 3-4 (2023): 397-411.

[59] Bodo Balazs, Natali Helberger, Kristina Irion, Frederik Zuiderveen Borgesius, Judith Moller, Bob van Velde de, Nadine Bol, Bram van Es, Claes de Vreese, "Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents" *Yale Journal Law & Technology*, 19 (2017): 171

[60] Rieke, Bogen, Robinson, "Public scrutiny of automated decisions: Early lessons and emerging methods."

5 | Non-Discrimination and the Risk-Based Approach in the AI Act 2024

The AI Act 2024/1689 represents a pivotal response to this regulatory gap. By introducing a risk-based approach, the Act aims to categorize AI systems based on their potential to cause harm, particularly in areas such as bias and discrimination.

The AI Act, officially Regulation (EU) 2024/1689,[61] is now the world's first comprehensive legal framework governing AI. Designed to promote trustworthy AI across Europe and secure the region's leadership in the global tech arena, this law builds on the original proposal from April 2021 and has now been adopted as binding legislation. A central feature of the AI Act is its risk-based approach, which classifies AI systems according to the potential harm they might cause. Each category comes with tailored requirements and safeguards:

Unacceptable risk: AI systems that might worsen biases or lead to discrimination are outright banned. This category covers, for example, systems used for social scoring or biometric categorisation that could unfairly target individuals based on factors like social behaviour or socioeconomic status.[62]

High risk: AI systems employed in critical areas, such as in managing essential infrastructure, job recruitment, border control, or loan approvals, fall into this category. Because these systems can have a major impact on society, they must satisfy strict "essential requirements" to ensure fairness and prevent inadvertent discrimination.[63]

Limited risk: although systems like chatbots might not seem harmful at first glance, the Act emphasizes transparency. Users must be clearly informed that they are interacting with an AI, ensuring they are aware of any potential biases in how the system operates.[64]

Minimal risk: for AI systems with minimal potential for discrimination, such as music or book recommendation engines, the regulatory requirements are lighter. Nonetheless, it remains important to consider how even these systems could reflect or perpetuate biases from the data they are built on.

[61] Regulation (EU) 2024/1689 of the European parliament and of the council of 13 June 2024 (AI Act 2024).

[62] Article 5 of AI Act 2024.

[63] Articles 6 and 7 of AI Act 2024.

[64] Article 8 of AI Act 2024.

5.1. Examining the Legal Framework for Non-Discrimination in the AI Act 2024/1689

The AI Act is the first truly transnational legal framework dedicated to promoting human-cantered, ethically developed AI. This law uniquely blends market oversight such as aspects of product liability with robust protection of fundamental rights. Two key issues emerge from this approach: first, the way the Act defines its risk categories plays a critical role in determining which technologies might foster discrimination; and second, while strict rules are imposed on high-risk systems, other AI applications face much lighter regulatory scrutiny.

5.1.1. The Challenge of Risk Categorization in AI Regulation

Categorization of the Act's risk-based model is both bold and contentious. By classifying AI systems according to their perceived threats, it aims to strike a balance between encouraging innovation and ensuring public safety. Yet, applying this framework in practice presents real challenges. For one, the rapid evolution of AI could quickly outdate the present categories or leave new, potentially risky systems, especially those with general-purpose functions, uncategorized. In response, the European Parliament has even been considering additional obligations for providers of foundational generative AI models. Another major concern is that the lines between some risk categories, particularly those leading to outright bans, can be quite blurry.[65] The current classification system combines factors like the potential to worsen biases with specific application contexts, sometimes resulting in exceptions that complicate enforcement. For example, Article 6 of Regulation (EU) 2024/1689 was designed to define high-risk AI systems by pinpointing sectors such as critical infrastructure, education, and employment, and by listing specific use cases in Annex III.

However, critics have pointed out that the criteria for what exactly makes an AI system "high-risk" are not exhaustively detailed, leaving room for interpretation.[66] Recent feedback from civil society groups highlights

[65] Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. "Taking AI risks seriously: a new assessment model for the AI Act" *AI & SOCIETY*, No. 5 (2024): 2493-2497.

[66] Isabel Kusche, "Possible harms of AI and the EU AI act: fundamental rights and risk" *Journal of Risk Research* (2024): 1-14.

that certain exemption conditions in Article 6 might allow developers too much leeway in deciding whether their systems qualify as high-risk, potentially undercutting the law's overall effectiveness. For instance, AI applications originally classified as high-risk, such as those used to monitor student behaviour, assess creditworthiness, screen job applicants, or determine eligibility for welfare benefits. Under the initial provisions,[67] developers and operators were required to ensure these systems were fair, transparent, and free from discriminatory biases. If the current loopholes persist, however, these safeguards might not be as effective as intended. This evolving framework reflects the balance between safeguarding individual rights and supporting technological progress, a balance that will continue to be refined as the AI landscape changes.

5.1.2. The Effectiveness of the Requirements on High-Risk AI Systems

The AI Act sets out mandatory requirements for high-risk AI systems, and while it recommends that as many AI systems as possible follow these rules, the focus is especially sharp on those deemed high-risk. These systems must implement robust risk management processes, maintain strict data handling and governance practices, compile detailed technical documentation, keep comprehensive records, adhere to transparency provisions, ensure meaningful human oversight, and meet stringent standards for accuracy, resilience, and cybersecurity.[68] At its core, these requirements aim to protect society from the broad risks of AI while specifically addressing issues of bias and discrimination.[69] Providers of high-risk systems must show that they have taken concrete steps at every stage, from development to deployment, to prevent biases. Traceability and explaining ability are essential, ensuring that fairness is built into every process. The Act also introduces Article 5, which outlines general principles that apply to all AI systems. These principles include human agency and oversight, technical robustness and safety, privacy and data governance, transparency, and particularly

[67] Articles 9 to 15 of AI Act 2024.

[68] Ahmad Khalil, Mohammad Bitar, S. Ananda Krishna Raj, "A New Era of Armed Conflict: The Role of State and Non-State Actors in Cyber Warfare with Special Reference to Russia-Ukraine War." *TalTech Journal of European Studies* 14, No. 2 (2024).

[69] Beatriz M. Cabrera, Luiz E. Luiz, João P. Teixeira, "The AI Act: Insights regarding its application and implications" *Procedia Computer Science*, 256 (2025): 230-237.

diversity, non-discrimination, and fairness.[70] Under these guidelines, AI systems must be developed and used in an inclusive way that promotes equal access, gender equality, and cultural diversity, all while avoiding discriminatory impacts and unfair biases as prohibited under Union or national law. Moreover, the Act mandates that AI systems be designed to be accurate, robust, safe, and secure. These standards are not only vital for general safety but also for preventing algorithmic discrimination that can occur due to inaccurate or biased data. To support this, the Act requires technical safeguards to protect against data poisoning or adversarial machine learning, which might otherwise lead to discriminatory outcomes.[71]

As these regulatory developments shape the broader AI landscape, it is crucial to assess their implications for specific sectors, such as higher education and research. While the AI Act provides a legal framework to address bias and discrimination in AI, its sectoral impact varies depending on risk classification.[72] Many AI systems in education,[73] including automated grading and research analysis tools, may not fall under the high-risk category, leaving potential gaps in regulatory oversight.[74] This highlights the need for continuous evaluation of AI applications in academic settings to ensure compliance with fairness and non-discrimination principles.[75]

Implications for Higher Education and Research The AI Act 2024 marks a significant step in regulating AI bias and discrimination. However, while the Act outlines strict requirements for high-risk AI systems, its direct impact on AI tools used in higher education and research remains ambiguous. Many AI-driven systems in academia, such as those used for

[70] Luca Deck, Jan-Laurin Müller, Conradin Braun, Domenique Zipperling, Niklas Kühl, *Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness*, (2024).

[71] Bert Heinrichs, "Discrimination in the age of AI" *AI & society*, No. 1 (2022): 143-154.

[72] Gabriel Bangura, *The European Union AI Act: Mitigating Discrimination In AI Systems*, (2024).

[73] Ahmad Khalil, S. Ananda Krishna Raj, "Deployment of autonomous weapon systems in the warfare: Addressing accountability gaps and reformulating international criminal law." *Balkan Social Science Review 23*, No. 23 (2024): 261-285.

[74] Ahmad Khalil, S. Ananda Krishna Raj, "Assessing the Legality of Autonomous Weapon Systems: An In-depth Examination of International Humanitarian Law Principles." *Law Reform 19*, No. 2 (2024): 372-392.

[75] Ahmad Khalil, Mohammad Bitar, S. Ananda Krishna Raj, "Navigating legal frontiers in cyber warfare: insights from the Russia-Ukraine conflict." *The Lawyer Quarterly 14*, No. 2 (2024).

admissions, grading, and research analysis, fall under "limited risk" categories, which require transparency but lack the strict oversight imposed on high-risk AI. As a result, the AI Act provides a foundation for addressing bias in AI tools used in education, but it does not fully resolve the concerns highlighted in this study. Future amendments and sector-specific regulations may be necessary to ensure that AI applications in education align with non-discrimination and fairness principles.

6 | Conclusion

In conclusion, addressing biases in AI systems, particularly in the context of higher education, necessitates a nuanced understanding of indirect discrimination. Indirect discrimination occurs when seemingly neutral practices or decisions disproportionately impact certain groups based on protected characteristics such as race, gender, or socioeconomic status. In higher education, AI algorithms used for tasks like student admissions or faculty hiring may inadvertently perpetuate indirect discrimination if they rely on historical data that reflects systemic biases. For example, if historical admission data favours applicants from privileged backgrounds, an AI admissions system trained on this data may perpetuate disparities by favouring similar candidates in the future. Legal frameworks must account for the complexities of indirect discrimination in AI systems. While laws typically prohibit direct discrimination based on protected characteristics, addressing indirect discrimination requires additional measures to identify and mitigate bias. This may involve implementing transparency measures to scrutinize AI decision-making processes and ensure that they do not inadvertently disadvantage certain groups. Collaborative efforts between AI developers, higher education institutions, and regulatory bodies are essential for addressing indirect discrimination. By working together to identify and address biases in AI algorithms, stakeholders can develop more equitable and inclusive systems that promote fair opportunities for all individuals.

Moving forward, continued research into methods for detecting and mitigating indirect discrimination in AI models is crucial. By incorporating diverse perspectives and experiences into the development and evaluation of AI systems, stakeholders can mitigate the risk of unintended biases and foster a more inclusive higher education environment that values fairness and equity for all.

As AI regulation continues to evolve, it is crucial to ensure that legal frameworks remain adaptable to emerging challenges. While the AI Act 2024/1689 represents a significant step towards mitigating bias and discrimination, its long-term effectiveness will depend on continuous assessment and refinement. The dynamic nature of AI technology requires ongoing dialogue between policymakers, educators, and AI developers to identify gaps in regulation and address new forms of algorithmic bias. Strengthening enforcement mechanisms, increasing transparency, and fostering ethical AI development will be key to ensuring that AI tools support fairness, inclusivity, and human rights in higher education and beyond.

Bibliography

Angwin Duncan N., Kamel Mellahi, Emanuel Gomes, Emmanuel Peter, "How communication approaches impact mergers and acquisitions outcomes" *The International Journal of Human Resource Management*, No. 20 (2016): 2370-2397.

Bagnall Anthony, Lines Jason, Bostrom Aaron, Large James, Keogh Eamonn, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances" *Data mining and knowledge discovery*, 31 (2017): 606-660.

Balazs Bodo, Natali Helberger, Kristina Irion, Frederik Zuiderveen Borgesius, Judith Moller, Bob van Velde de, Nadine Bol, Bram van Es, Claes de Vreese, "Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents" *Yale Journal Law & Technology*, 19 (2017): 133-180.

Barocas Solon, Selbst D. Andrew, "Big data's disparate impact" *California Law Review*, 104 (2016): 671-732.

Bengio Yoshua, Ducharme Réjean, Vincent Pascal, "A neural probabilistic language model" *Advances in neural information processing systems*, 13 (2000).

Bitar Mohammad, Benarji Chakka, "Drone attacks during armed conflict: quest for legality and regulation." *International Journal of Intellectual Property Management 13*, No. 3-4 (2023): 397-411.

Bolukbasi Tolga, Chang Kai-Wei, Zou Y James, Saligrama Venkatesh, Kalai T. Adam, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings" *Advances in neural information processing systems*, 29 (2016).

Bordia Shikha, Samuel R. Bowman, "Identifying and reducing gender bias in word-level language models" *arXiv preprint arXiv:1904.03035* (2019).

Brown Ian, Christopher T. Marsden, *Regulating code: Good governance and better regulation in the information age*. Cambridge: MIT Press, 2023.

Brusseau James, "AI human impact: toward a model for ethical investing in AI-intensive companies" *Journal of Sustainable Finance & Investment*, No. 2 (2023): 1030-1057.

Bryson John, Alessandro Sancino, John Benington, Eva Sørensen, "Towards a multi-actor theory of public value co-creation" *Public Management Review*, No. 5 (2017): 640-654.

Bubeck Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee et al. "Sparks of artificial general intelligence: Early experiments with gpt-4" *arXiv preprint arXiv:2303.12712*, (2023).

Buolamwini Joy, Gebru Timnit, "Gender shades: Intersectional accuracy disparities in commercial gender classification" *Proceedings of the 1st Conference on fairness, accountability and transparency*, 81 (2018): 77-91.

Caliskan Aylin, Joanna Bryson, Narayanan Arvind, "Semantics derived automatically from language corpora contain human-like biases" *Science*, No. 6334 (2017): 183-186.

Chouldechova Alexandra, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments" *Big Data*, No. 2 (2017): 153-163.

Conneau Alexis, Guillaume Lample, Marc Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou, "Word translation without parallel data" *arXiv preprint arXiv:1710.04087* (2017).

Daelman Charline, Katerina Yordanova, "AI through a human rights lens. The role of human rights in fulfilling AI's potential", [in:] *Artificial intelligence and the law*. 135-168. Cambriodge: Intersentia, 2023.

Devlin Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "Pre-training of deep bidirectional transformers for language understanding" *arXiv preprint arXiv:1810.04805* (2018).

Dixon Lucas, Li John, Sorensen Jeffrey, Thain Nithum, Vasserman Lucy, "Measuring and mitigating unintended bias in text classification", [in:] *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018.

Doshi-Velez Finale, Been Kim, "Towards a rigorous science of interpretable machine learning" *arXiv preprint arXiv:1702.08608*, (2017).

Ferrara Emilio, “The history of digital spam” *Communications of the ACM*, No. 8 (2019): 82-91.

Ferrara Emilio, “GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models” *Journal of Computational Social Science*, (2024): 1-21.

Hamilton James D., “Why you should never use the Hodrick-Prescott filter” *Review of Economics and Statistics*, No. 5 (2018): 831-843.

Holm Sune Hannibal, Kasper Lippert-Rasmussen, “Discrimination, Fairness, and the Use of Algorithms” *Res Publica*, No. 2 (2023): 177-183.

Jawaid Shaikh Sonia, “Artificially intelligent, interactive, and assistive machines: A definitional framework for intelligent assistants” *International Journal of Human–Computer Interaction*, No. 4 (2023): 776-789.

Khaitan Tarunabh, *A theory of discrimination law*. Oxford: Oxford Univeristy Press, 2015.

Khalil Ahmad, Bitar Mohammad, Raj Ananda Krishna S., “A New Era of Armed Conflict: The Role of State and Non-State Actors in Cyber Warfare with Special Reference to Russia-Ukraine War.” *TalTech Journal of European Studies 14*, No. 2 (2024).

Khalil Ahmad, Bitar Mohammad, Raj Ananda Krishna S., “Navigating legal frontiers in cyber warfare: insights from the Russia-Ukraine conflict.” *The Lawyer Quarterly 14*, No. 2 (2024).

Khalil Ahmad, Raj Ananda Krishna S., “Assessing the Legality of Autonomous Weapon Systems: An In-depth Examination of International Humanitarian Law Principles.” *Law Reform 19*, No. 2 (2024): 372-392.

Khalil Ahmad, Raj Ananda Krishna S., “Challenges to the Principle of Distinction in Cyber Warfare Navigating International Humanitarian Law Compliance: The Principle of Distinction in Cyber Warfare.” *Prawo i więź 2* (2024): 109-131.

Khalil Ahmad, Raj Ananda Krishna S., “Deployment of autonomous weapon systems in the warfare: Addressing accountability gaps and reformulating international criminal law.” *Balkan Social Science Review 23*, No. 23 (2024): 261-285.

Kleinberg Jon, Sendhil Mullainathan, Raghavan Manish, “Inherent trade-offs in the fair determination of risk scores” *arXiv preprint arXiv:1609.05807*, (2016).

Koops Bert-Jaap, “Should ICT regulation be technology-neutral?”, [in:] *Starting Points for ICT Regulation – Deconstructing Prevalent Policy One-liners*, ed. Bert Jaap Koops et al. Den Haag: Asser, 2006.

Kristian Lum, William Isaac, “To predict and serve?” *Significance*, No. 5 (2016): 14-19.

Kroll David S., Harry Reyes Nieva, Arthur J. Barsky, Jeffrey A. Linder, “Benzodiazepines are prescribed more frequently to patients already at risk for benzodiazepine-related adverse events in primary care" *Journal of General Internal Medicine*, 31 (2016): 1027-1034.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado S Greg and Dean Jeff, "Distributed representations of words and phrases and their compositionality" *Advances in neural information processing systems* (2013).

O'Reilly-Shah Vikas, Gentry R. Katherine, Cleve Van Wil, Kendale M. Samir, Jabaley S. Craig, and Long R Dustin, "The COVID-19 pandemic highlights shortcomings in US health care informatics infrastructure: a call to action" *Anesthesia & Analgesia*, No. 2 (2020): 340-344.

Omer Tene, Jules Polonetsky, "Taming the Golem: Challenges of ethical algorithmic decision-making" *North Carolina Journal of Law & Technology*, No. 1 (2017): 125-173.

Pasquale Frank, *The black box society: The secret algorithms that control money and information*. Cambridge: Harvard University Press, 2015.

Peck Tabitha C., Seinfeld Sofia, Aglioti M. Salvatore, Mel Slater, "Putting yourself in the skin of a black avatar reduces implicit racial bias" *Consciousness and Cognition*, No. 3 (2013): 779-787.

Popenici Stefan A.D., Kerr Sharon, "Exploring the impact of artificial intelligence on teaching and learning in higher education" *Research and practice in technology enhanced learning*, No. 1 (2017).

Prates Marcelo, Pedro H. Avelar, Luís C. Lamb, "Assessing gender bias in machine translation: a case study with google translate" *Neural Computing and Applications*, 32 (2020): 6363-6381.

Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, Sutskever Ilya, "Language models are unsupervised multitask learners" *OpenAI blog*, No. 8 (2019).

Rieke Aaron, Miranda Bogen, David G. Robins, *Public scrutiny of automated decisions: Early lessons and emerging methods*. Upturn and Omidyar Network, 2018.

Rosamond Mitchell, Florence Myles, Emma Marsden, *Second language learning theories*. London: Routledge, 2019.

Schick Timo, Hinrich Schütze, "It's not just size that matters: Small language models are also few-shot learners" *arXiv preprint arXiv:2009.07118* (2020).

Sirsendu Jana, Michael R. Heaven, Charles B. Stauft, Tony T. Wang, Matthew C. Williams, Felice D'Agnillo, Abdu I. Alayash, "HIF-1α-Dependent metabolic reprogramming, oxidative stress, and bioenergetic dysfunction in SARS-CoV-2-infected hamsters" *International Journal of Molecular Sciences*, No. 1 (2022).

Yang Kai-Cheng, Ferrara Emilio, Menczer Filippo, "Botometer 101: Social bot practicum for computational social scientists" *Journal of Computational Social Science*, No. 2 (2022): 1511-1528.

Zellers Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi, “Defending against neural fake news” *Advances in neural information processing systems* (2019).

Zuiderveen Frederik Borgesius, *Discrimination, artificial intelligence, and algorithmic decision-making*. Council of Europe, Directorate General of Democracy, 2018.