Üçüncü Sektör Sosyal Ekonomi Dergisi
Third Sector Social Economic Review
58(4) 2023, 3335-3345
doi: 10.15659/3.sektor-sosyal-ekonomi.23.11.2302

Review Article

Algorithmic Bias as a Sociological Issue: How Big is the Challenge for 'AI for Social Good'?[1]

Sosyolojik Bir Sorun Olarak Algoritmik Yanlılık: 'Sosyal Fayda için Yapay Zekâ'nın Zorlukları Ne Kadar Büyük?

Aysu KES ERKUL
Doç.Dr., Ankara Bilim Üniversitesi
Siyaset Bilimi ve Kamu Yönetimi Bölümü
aysu.erkul@ankarabilim.edu.tr
https://orcid.org/0000-0001-6283-0606

Makale Geliş Tarihi	Makale Kabul Tarihi
07.11.2023	**29.11.2023**

Abstract

Artificial Intelligence (AI) undeniably has been the center of attention when it comes to technology and social change. The ever- increasing penetration of AI applications to various segments of social and daily life is causing substantial positive and negative social impact. One of the major reasons for the negative or unwanted social impact of AI is originating from AI bias rooted in the development process of AI itself, namely data collection and handling as well as the design of algorithms. On the other hand, recent developments in AI and related technologies raised hope for positive social impact as well as finding solutions for the social and environmental problems of humanity. Artificial Intelligence for Social Good (AI4SG) concept has been developed with expectations to use state-of-the-art AI tools and applications for the benefit of whole society. However, AI bias is a significant challenge towards achieving the goals set by Artificial Intelligence for Social Good projects. The alleviation of AI bias with reference to ethical and human-oriented guidelines may ensure successful results. In this regard, this paper discusses the challenges of AI toward achieving AI4SG goals.

Key Words: AI bias, Algorithmic bias, AI4SG, decision support systems, datafication

Öz

Yapay Zekâ (AI), teknoloji ve sosyal değişim söz konusu olduğunda inkâr edilemez bir şekilde ilgi odağı olmuştur. Yapay zekâ uygulamalarının toplumsal ve günlük yaşamın çeşitli alanlarında giderek yaygınlaşması, kayda değer biçimde olumlu ve olumsuz toplumsal etkilere neden olmaktadır. Yapay zekanın olumsuz veya istenmeyen sosyal etkilerinin ana nedenlerinden biri, yapay zekanın geliştirilmesi sürecinden, yani veri toplama ve işlemenin yanı sıra algoritma tasarımından kaynaklanan yapay zekâ önyargısından ileri gelmektedir. Öte yandan yapay zekâ ve ilgili teknolojilerdeki son gelişmeler, insanlığın sosyal ve çevresel sorunlarına çözüm bulunmasının yanı sıra olumlu sosyal etki yaratma

[1] 'A previous narrow version of this paper was presented at 'XX. ISA World Congress of Sociology, between June 25th - July 1st, 2023, Melbourne, Australia '.
'Bu makalenin dar kapsamlı önceki versiyonu 25 Haziran- 1 Temmuz , 2023 tarihklerinde Melbourne, Avusturalya'da yapılan   '10. Avrupa Sosyoloji Kongresi'nde sunulmuştur.

Önerilen Atıf /Suggested Citation
Kes Erkul, A., 2023, Algorithmic Bias as a Sociological Issue: How Big is the Challenge for 'AI for Social Good'?, *Üçüncü Sektör Sosyal Ekonomi Dergisi*, 58(4), 3335-3345.

umutlarını da artırmıştır. Sosyal Fayda için Yapay Zekâ (AI4SG) kavramı da en son teknolojiye sahip yapay zekâ araç ve uygulamalarının tüm toplumun yararına kullanılması beklentisiyle geliştirilmiştir. Ancak yapay zekâ yanlılığı, Sosyal Fayda için Yapay Zekâ projelerinin belirlediği hedeflere ulaşması bakımından önemli bir zorluk teşkil etmektedir. Yapay zekâ yanlılığının etik ve insan odaklı yönergelere referansla hafifletilmesi başarılı sonuçlar elde edilmesini sağlayabilir. Bu bağlamda, bu makale 'Sosyal Fayda için Yapay Zekâ' hedefine ulaşmada yapay zekanın karşılaştığı zorlukları tartışmaktadır.

Anahtar Kelimeler: Yapay Zekâ yanlılığı, Algoritma Yanlılığı, Sosyal Fayda için Yapay Zekâ, karar destek sistemleri, verileştirme

1.Introduction

The popularity of technological solutions to social, economic, and environmental problems is on the rise, especially since the outbreak of COVID-19 pandemic. The use of platforms, IoT applications, web-based services, distance learning and working technologies became more and more widespread in the process. Additionally, the increase in the climate change concerns turned the attention to technological innovation as hope for solutions. Meanwhile, availability of big data with the increased use of digital devices, applications and IoT paved the way for the acceleration of digitalization in all areas. Also, the technologies that are based on big data keep improving thanks to data science and machine learning. As a result of all these developments, Artificial Intelligence became the focus of innovation in recent years. Now we are at a point where 'innovation is twofold'. On the one hand, there is AI and everything else on the other. In other words, the pace of AI based innovations are dominating the technological disruption today. The widening use of AI in many sectors and processes such as marketing, finance, health services, policing etc. triggers various questions regarding future of economy and government as well as the society in general. Many businesses and public institutions are using AI based decision support systems which are supposedly helping humans to make efficient decisions. These systems are almost always coming with their social impact. The recent debate on that subject is focused on the unpredictability and biased outcomes of AI and on finding ways to regulate the utilization (together with the development) of AI applications. Therefore, it is essentially important to observe and critically analyze the social impact of AI in general. Particularly, the decisions made by algorithms must be meticulously studied to understand the process as a social phenomenon that is beyond technological innovation. AI and ML based decision making may offer great help for better policy making(Fountain, 2022). However, many case studies and research regarding the outcomes and impact of machine assisted decision making reveals the fact that the promise of AI is not realized flawlessly. At many stages of the process, there are potential issues that make the products debatable. In order to clarify the interrelatedness of artificial intelligence and society in general, it is important to frame the sociological aspects of AI. Thereby, it can be possible to identify the existing and potential problem areas.

2. What is Sociological about Artificial Intelligence?

Artificial Intelligence refers to smart machines that manipulate language, knowledge, and relations by learning the patterns in human generated data (Airoldi, 2021). In other words, machines with AI process the information that is fed to them to make decisions and tell us results of specific analysis asked from them. Additionally, AI as a computer-based system that recognizes patterns in data can also apply these patterns to new data for specific task (Livingston, 2020). The algorithm on the other hand determines the method of processing. Therefore, data and algorithm that are constitute human- made contributions to the system are the major components of AI. On the other hand, the major factor behind the dramatic change through AI, is the extraordinary disposal human behavioral data (Lepri et al., 2016). An algorithm in that context, is technically (and simply) a set of instructions used by computers to solve a problem, make decisions or predictions (Boscoe, 2019). These algorithms can identify, select, and determine significance 'beyond human decision-making which creates a new kind of decision optimization' through both the design of the algorithms and the data from which they are based on (Lepri et al., 2016). First of all, the age of datafication roughly starting with the introduction of internet and search engines, followed by early recommendation systems turned world wide web into a huge platform that paved the way for 'algorithmic ordering of economy and society'(Airoldi, 2021). As the humans started to rely on data and machine learning to make decisions, the nature of social dynamics also started to change. In the process, data became 'big data' as a byproduct of various types of human activity with the ever-increasing penetration of smart phones, sensors, social media etc.(Coulton, Goerge, Putnam-Hornstein, & De Haan, 2015). The volume, velocity, and variety (three v's of data) started to increase which made more AI based decision making possible. This in turn created elevated impact. It is widely accepted that big data has the potential of accelerating scientific advancement and innovation as well as creating positive social change(Drosou, Jagadish, Pitoura, & Stoyanovich, 2017)

To better visualize the sociological nature of AI based decision support systems it is possible to break it down. Artificial intelligence consists of three main components in their making, just like any other AI application, such as computer vision/image processing and generative AI. Each of these components are related to 'social' and 'societal' in many ways. Accordingly, three components of AI that make it a sociological issue are data, design, and algorithm.

First of all, data is social. Machine Learning, as the technical backbone of AI, explains and constitutes the interaction between the machine and the social environment since it refers to the process of making predictions or classifications based on the training data (Boscoe, 2019). Existing or real-time data used to train the machines is the input from social environment which feed the output generated by the algorithmic process. Today 'almost everything can be transformed into machine-readable data' such as images, documents, sounds, movements, and behaviors (Airoldi, 2021). However, the design of algorithm and the AI application is more likely to determine the 'nature' of the outcome in terms of whether it produces objective and neutral outcomes.

Secondly, AI design is the process through which the ends, namely the data and application meet. Design of an AI product draws the outlines of the ways in which these components are put together, so that technical and social merge into an application. Through AI design, the data is given meaning in a way to capture patterns and clusters (Airoldi, 2021; Kaufmann et al., 2019). Given the interwoven nature of human and data in design, AI is not just a machine but a social agent in practice. This social agency is most evident in the societal impact of decisions made by the machine as well as in the biased outcomes and can be understood and studied as 'behavior' since the output of AI is shaped by algorithms and the social environments that they interact with (Rahwan et al., 2019). Moreover, the impact of this behavior is unprecedented. The use of data to support decision making has the potential of affecting various aspects large groups of people in many ways (Barocas et al., 2011). Looking at the three components as extensions of human intentions, as well as at their combined social impact, we can clearly comprehend the 'sociology' behind it. Thus, such behavior and/or social agency is a legitimate object of sociological research (Airoldi, 2021).

The discussion about societal impact of algorithms and AI are not new yet intensifying (Joyce et al., 2021). Algorithms that drive AI, are instruments of power in the sense that they make decisions about the lives of people, and they draw 'digital boundaries' among social groups through these decisions (Airoldi, 2021). They function as filters and mechanisms that define social processes through which actual social consequences occur. Therefore, the term 'sociotechnical system' is utilized to refer to the ways in which AI is intertwined with values, institutional practices, and inequalities (Joyce et al., 2021). This perspective also paves the way to the sociological analysis of AI design and use.

As AI penetrates various domains of life, the social impact becomes more observable. One of the most mentioned examples of this trend is the algorithms used for news- ranking and social media flows, since they heavily shape the content that the users see and interact with (Rahwan et al., 2019). Even basic Search Engine Optimization (SEO) algorithms are now shaping and altering our online experience all together. AI applications in financial sector are also a well- known example since they regulate the intersections by predicting risks and proposing investments. Labor market is also being reshaped by use of AI technology since the decisions about employment processes are made by algorithms. AI can be life-changing when AI is used to decide whether someone is eligible for a loan or a job (Boscoe, 2019). Many other use cases and examples are available regarding various parts of life.

3. AI Bias: Definition and Consequences

Sociologists, as well as other social scientists are skeptical about AI and their reservations are sparking criticisms about the use of AI in/for social purposes. The biased decisions made by AI algorithms are one reason for the critical discussion to arise. On the other hand, major tech companies as well as governmental institutions around the world are investing in and funding AI research and development

(Joyce et al., 2021). Therefore, the resource allocation to develop algorithmic solutions for societal issues becomes a topic of debate. This trend also turns the debate into a political one at some point. On the other hand, the technical complexity of AI may conceal the actual processes leading to specific results for social scientists to see, but examining the logic behind the AI processes can help to find the root causes (Boscoe, 2019).

In keeping with the purpose of this paper, algorithmic bias will be defined in detail, with reference to its several aspects. In its simplest form, bias can be defined as a deviation from the standard which may help to show the statistical patterns in the data (Ferrer et al., 2020). In data science and especially in machine learning, even simple biases alter the output significantly. Hence, the diagnosis of the root cause of the bias is essential to overcome this issue.

As mentioned above, algorithmic bias is a general concept that refers to all kinds of bias-related problems in machine learning which can be originating from different fragments of the process, so it is believed to be helpful to break it down. The causes of bias can be both technical and social: the biases of the designers and data can be embedded in the algorithm, and the use of AI can reproduce bias already existing in a population (Livingston, 2020). The way in which the data handled by AI designers and data scientists is crucial here since optimizing the outcome may lead to bias against the already disadvantaged groups included in the data set (Barocas et al., 2011). In that context, it is possible and necessary to discuss the root causes of bias under two categories, namely human- related bias and data related- bias.

3.1. Human- related bias: The bias resulting from human intentions and interventions in the process of AI design and data -handling can be framed as human- related bias. One form of this bias is related with the design of specific AI. People behind the design attribute value to data. This is the human contribution to machine learning by giving meaning to various types of raw data. Developers or AI designers may assume that the data they use to train the machine is neutral, representative, and truthful in terms of reflecting with the social realm (Joyce et al., 2021). Since the algorithm and the AI application is designed on the basis of this assumption, the outcome will not go beyond embedded discrimination and bias. The quality and effectiveness of algorithm cannot change that, and the AI application may look flawless. In fact, the opaqueness and assumed neutrality causes asymmetry and lack of transparency between algorithms and beneficiaries (Lepri et al., 2016). The developers may not be aware of the actual social meaning of certain data such as zip codes being an indicator of social status (Joyce et al., 2021). AI design that is shaped by human values may lead to bad results, if values are not selected properly (Floridi, Cowls, King, & Taddeo, 2020). There are many studies on such design processes that give unwanted results. Design, implementation, and use (all of which are human aspects of AI already in use are) producing inaccurate predictions and harmful decisions for certain social groups (Livingston, 2020).

Additionally, another major sociological aspect of AI design stems again from the human designers as social beings. The designers who are responsible for the development and writing of algorithms are individuals who make choices and set the goals for the AI application. Therefore, the sociological background of people is inscribed in the design of the final product and in the outputs in turn (Airoldi, 2021). For example, the people in tech sector are still dominantly men with similar educational backgrounds, interests and lifestyles which may cause the algorithm to be racist, sexist and /or classist (Airoldi, 2021). However, data that is not collected for a specific purpose, necessitate careful cleaning and validation by domain experts of relevant topic (Coulton et al., 2015)

3.2. Data- related Bias: AI is data- intensive. Aside from bias that is being transported to the AI through assumptions and processes mentioned above, the data itself is a point of concern from a social perspective. Training data and real-time data used in ML and AI involve different problems. Training data used in ML is most of the time loaded with existing prejudices which are transferred to AI through the process (Ferrer et al., 2020). For example, if the training data is underrepresenting women, then the AI will produce gender biased outputs. Data can 'reflect widespread, persistent societal-level biases' such as gender discrimination in recruitment (Fountain, 2022). This type of bias can also be approached from the perspective of 'diversity' or the lack of it, as both an ethical and practical issue (Drosou et al., 2017).

Collection of real- time data to feed the machine to produce precise decisions can also be problematic since the recorded data tend to be inherently prejudiced. There are several reasons for that. For example, willingness of people to share data is not equally distributed due to various reasons such as changing trust to applications and institutions that are collecting the data (Kontokosta & Hong, 2020). Individuals may refrain from using certain apps or sharing personal data of any kind with a concern about privacy or opaqueness of data processing.

Digital divide can be another reason for certain groups of people to be underrepresented in the data set. Technology acceptance makes a significance difference among people in terms of using digital services, especially those require data sharing (Kontokosta & Hong, 2020; Shin, Kim, & Chun, 2021). The digital literacy is also another issue that effects the effective use of technology in general, including the ones that are collecting data (Shin et al., 2021).

An additional aspect of AI that is problematic from the perspective of bias and decision making is the self- learning AI, a.k.a. neural networks, or artificial neural networks (ANNs). In simplest terms, artificial neural networks refer to the connections between variables in the algorithm. When the data is flawed in any of the ways mentioned above, the ANN or the self- learning algorithm may give flawed decisions even if the risks for biased decisions are excluded by the algorithm. Especially in supervised learning that involves the introduction of a training data set, the machine learns human judgements and carry them into ANN (Umbrello & van de Poel, 2021).

4. AI for Social Good- Beyond Data and Algorithm

AI for social good (AI4SG) is a term basically used to refer to the AI projects that are noncommercial and aiming at producing socially beneficial outcomes(Cowls et al., 2021). Especially with the pressing problems of humanity such as sustainability and crisis like pandemics AI and related technologies are prioritized to offer social solutions(Holzmeyer, 2021; Magalhães & Couldry, 2021). The encounter of seemingly unrelated fields of AI and social problems is basically about the presumption that datafication within social service, education, health, and many other sectors can facilitate a better understanding of social problems as well as finding effective solutions (Coulton et al., 2015). However, this framework must be defined in more detail to understand the scope and aim of such projects. Since artificial intelligence and social good are two concepts with their own complexities, the joined definition needs description. To begin with, both AI and social good have their specific links with what is social. As argued above, the data used in AI processes is inherently social which refers to the origins and nature of the material. From the perspective of social good, on the other hand, the meaning of social is more extensive since it implies the scope of impact. In other words, the ‘social’ in social good is more comprehensive. Therefore, the term ‘social good’ itself is hard to delimit as it is loaded with value. Obviously, ‘non-profit’ or ‘social benefit’ does not always entail ‘social good’. The term in general assumes a generalized benefit for the society as a whole.

Artificial intelligence relies of data. However, AI for social good is beyond Data Science and Data Science for Social Good. Firstly, AI may never need supervision and even training. Secondly, AI for Social Good augments previous technologies together with data science (Cowls et al., 2021). AI4SG is an extension of ‘technology for good’ initiatives (Holzmeyer, 2021). Therefore, processing data regarding a social issue may not necessarily mean AI4SG.

Many scholars proposed various sets of principles to combine AI and social good in a relatively tangible way. For instance, to simplify the framework, three basic requirements are suggested for AI4SG projects as preconditions; (1) either prevent, mitigate and/or resolve problems that harm human life and/or the welfare of the natural world or (2) facilitate socially desirable or environmentally sustainable developments and (3) should not cause new forms of harm and/or expand existing disparities and inequalities(Cowls et al., 2021).

5. Critical Views

Aside from AI bias, AI4SG as an approach to solving many social, economic, and environmental problems of our times is criticized from various vantage points. The focus of this paper is AI bias, but some other criticisms are worth briefly mentioning to take a look at the tension between AI and social good. For instance, data privacy and security are discussed in that context. The datafication of personal information and utilization of data for humanitarian purposes may increase vulnerabilities when it comes to already disadvantaged groups such as refugees or disaster victims (Coulton et al., 2015; Magalhães & Couldry, 2021).

Another point of criticism regarding AI4SG is about excessive reliance on big data. If a specific problem is affecting a small section of the population or the affected group is underrepresented in the data set, the problem may not be noticed (Coulton et al., 2015).

AI4SG projects are also disapproved for their tendency to draw attention away from the root causes of social and environmental problems. Tech companies are argued to be putting their AI abilities to the fore (Holzmeyer, 2021). The concerns about consent and access to data and/or machine learning follows (Andrejevic, 2014; Magalhães & Couldry, 2021).

6. Suggested Solutions in a Nutshell

The ongoing debate about bias in AI applications and the way in which bias (as well as other issues in AI) effect 'social good' defectively produced discussions about the potential remedies, too. On the technical side, it is possible to talk about creating better algorithms, but even such a solution requires a fresh perspective about the nature of the problem, namely the social/sociological characteristic of the technology.

Transparency was also suggested as a conceptual solution with many practical prerequisites including data collection and design stages (Boscoe, 2019). Human supervision and utilization of domain expertise are considered key to overcome roadblocks in the process. Procedures for human supervision must be strong enough to minimize the data – related bias and even human related bias in AI (Taddeo & Floridi, 2018). Inclusive AI is another concept/principle that is suggested as a solution to overcome bias (and discrimination) in artificial intelligence. A brief definition of the term refers to Inclusive AI as aiming heterogeneity through responses of generative AI (Arumugam, Dong, & Van Roy, 2022). Inclusiveness is mostly used together with other related concepts such as equity and fairness to address ethical principles that are necessary to overcome bias and discrimination in the process (Cachat-Rosset & Klarsfeld, 2023).

In recent years, governments, NGOs, and various other organizations released guidelines and regulatory documents for artificial intelligence. These documents increasingly address ethical and social dimensions of AI as a growing issue. Therefore, bias and discrimination are taking a substantial part in guidelines. High- Level Expert Group on Artificial Intelligence (ALTAI) published an assessment list and tool for self-assessment in 2020. In the document, seven requirements are listed as: (1) Human Agency and Oversight, (2) Technical Robustness and Safety, (3) Privacy and Data Governance, (4) Transparency, (5) Diversity, Non-discrimination, and Fairness (6) Societal and Environmental Well-being, and (7) Accountability(High-Level Expert Group on Artificial Intelligence, 2020). All of these items actually address principles of Artificial intelligence for Social good as well as the social and ethical concerns mentioned above, including bias and discrimination. Also, this list of requirements implies more human involvement and intervention to the AI processes from design to use as a general approach to making AI trustworthy and reliable.

7. Conclusion

Artificial intelligence is the most debated innovation of our times with both its technological potential and growing social impact. Unlike many previous innovations, AI is penetrating into almost all sectors of life from health to education and from production to communication. The unparalleled abilities of it also carries potential for solving social and environmental problems. If utilized right, AI can help generate social good for general society in many ways. The endeavor to use AI as a tool for human development and improvement of life conditions is executed under the concept of Artificial Intelligence for Social Good. However, both AI and Social Good have their limitations and ambiguities. Many experiences with AI tools show that the outcomes can be socially harmful due to bias and discrimination (re)generated by the several stages of the process from data collection to algorithm. On the other hand, social good can be an overgeneralization. In that context, elimination of bias in AI is the most crucial stage towards achieving desired goals through AI. Regulating human intervention to AI process through ethical guidelines and regulations are the key to obtain unbiased AI outcomes. Technical measures and improvement of system to tackle the challenge of bias can only be meaningful with empowering human aspect of AI. It should be kept in mind that like all innovations, AI should serve human and social benefits. In the light of the discussions above, the future research should focus on detailed analysis of human- related root causes of AI bias and universal guidelines and overarching values, which requires a comprehensive discussion with all the stakeholders and beneficiaries.

References

Airoldi, M. (2021). Machine Habitus: Toward a Sociology of Algorithms. John Wiley&Sons.

Andrejevic, M. (2014). The Big Data Divide. In International Journal of Communication (Vol. 8). Retrieved from http://ijoc.org.

Arumugam, D., Dong, S., & Van Roy, B. (2022). Inclusive Artificial Intelligence. Retrieved from http://arxiv.org/abs/2212.12633

Barocas, S., Selbst, A. D., Bambauer, J., Bedoya, A., Blumenthal, M., Citron, D., … Vladeck, D. (2011). Electronic Privacy Information Center. J.D. https://doi.org/10.15779/Z38BG31

Boscoe, B. (2019). Creating Transparency in Algorithmic Processes. Delphi - Interdisciplinary Review of Emerging Technologies, 2(1), 12–22. https://doi.org/10.21552/delphi/2019/1/5

Cachat-Rosset, G., & Klarsfeld, A. (2023). Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines. Applied Artificial Intelligence, 37(1). https://doi.org/10.1080/08839514.2023.2176618

Coulton, C. J., Goerge, R., Putnam-Hornstein, E., & De Haan, B. (2015). American Academy of Social Work and Social Welfare Harnessing Big Data for Social Good: A Grand Challenge for Social Work Grand Challenges for Social Work Initiative Grand Challenge 12: Harness Digital Technology for Social Good.

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021, February 1). A definition, benchmark and database of AI for social good initiatives. Nature Machine Intelligence, Vol. 3, pp. 111–115. Nature Research. https://doi.org/10.1038/s42256-021-00296-0

Drosou, M., Jagadish, H. V., Pitoura, E., & Stoyanovich, J. (2017, June 1). Diversity in Big Data: A Review. Big Data, Vol. 5, pp. 73–84. Mary Ann Liebert Inc. https://doi.org/10.1089/big.2016.0054

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2020). Bias and Discrimination in AI: a cross-disciplinary perspective. https://doi.org/10.1109/MTS.2021.3056293

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. Science and Engineering Ethics, 26(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Fountain, J. E. (2022). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. Government Information Quarterly, 39(2). https://doi.org/10.1016/j.giq.2021.101645

High-Level Expert Group on Artificial Intelligence. (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI). Brussels: European Commission .

Holzmeyer, C. (2021). Beyond ‘AI for Social Good’ (AI4SG): social transformations—not tech-fixes—for health equity. Interdisciplinary Science Reviews, 46(1–2), 94–125. https://doi.org/10.1080/03080188.2020.1840221

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., … Shestakofsky, B. (2021). Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change. Socius, 7. https://doi.org/10.1177/2378023121999581

Kaufmann, M., Egbert, S., & Leese, M. (2019). Predictive policing and the politics of patterns. British Journal of Criminology, 59(3), 674–692. https://doi.org/10.1093/bjc/azy060

Kontokosta, C. E., & Hong, B. (2020). Bias in Smart City Governance: How Socio-Spatial Disparities in 311 Complaint Behavior Impact the Fairness of Data-Driven Decisions.

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2016). The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good. Retrieved from http://arxiv.org/abs/1612.00323

Livingston, M. (2020). Preventing Racial Bias in Federal AI. Journal of Science Policy & Governance, 16(02). https://doi.org/10.38126/jspg160205

Magalhães, J. C., & Couldry, N. (2021). Giving by Taking Away: Big Tech, Data Colonialism, and the Reconfiguration of Social Good. In International Journal of Communication (Vol. 15). Retrieved from http://ijoc.org.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., … Wellman, M. (2019, April 25). Machine behaviour. Nature, Vol. 568, pp. 477–486. Nature Publishing Group. https://doi.org/10.1038/s41586-019-1138-y

Shin, S. Y., Kim, D., & Chun, S. A. (2021). Digital divide in advanced smart city innovations. Sustainability (Switzerland), 13(7). https://doi.org/10.3390/su13074076

Taddeo, M., & Floridi, L. (2018, August 24). How AI can be a force for good. Science, Vol. 361, pp. 751–752. American Association for the Advancement of Science. https://doi.org/10.1126/science.aat5991

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. AI and Ethics, 1(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3

Derleme Makale

Sosyolojik Bir Sorun Olarak Algoritmik Yanlılık: 'Sosyal Fayda için Yapay Zekâ'nın Zorlukları Ne Kadar Büyük?

Algorithmic Bias as a Sociological Issue: How Big is the Challenge for 'AI for Social Good'?

Aysu KES ERKUL

Doç.Dr., Ankara Bilim Üniversitesi

Siyaset Bilimi ve Kamu Yönetimi Bölümü

aysu.erkul@ankarabilim.edu.tr

https://orcid.org/0000-0001-6283-0606

Genişletilmiş Özet

Özellikle COVID-19 salgınının patlak vermesinden bu yana sosyal, ekonomik ve çevresel sorunlara yönelik çözümlerin geliştirilmesi için teknolojiye olan ilgi artmıştır. Diğer yandan dijitalleşmenin de hız kazanmasıyla büyük veriye dayalı teknolojiler, veri bilimi ve makine öğrenimi sayesinde önemli gelişme göstermiştir. Tüm bu gelişmelerin sonucunda yapay zekâ son yıllarda inovasyonun odağı haline gelmiştir. Yapay zekanın pazarlama, finans, sağlık hizmetleri, eğitim vb. birçok sektör ve süreçte kullanımının yaygınlaşması, toplumun geleceğine ilişkin çeşitli soruları da beraberinde getirmektedir. Yapay zekâ ve toplum arasındaki ilişkiyi açıklığa kavuşturmak için yapay zekanın sosyolojik yönlerini tanımlamak önemlidir. Böylece mevcut ve potansiyel sorun alanlarının tespit edilmesi mümkün olabilecektir.

Yapay zekâ, insan tarafından üretilen verilerdeki kalıpları öğrenerek dili, bilgiyi ve ilişkileri yönlendiren akıllı makineleri ifade eder (Airoldi, 2021). Verilerdeki kalıpları tanıyan bilgisayar tabanlı bir sistem olan yapay zekâ, bu kalıpları belirli bir görev için yeni verilere de uygulayabilir (Livingston, 2020). Yapay zekâ alanında yaşanan dramatik ilerlemenin ardındaki en önemli faktör, insan davranış verilerinin yaygın olarak kullanılabilir hale gelmiş olmasıdır (Lepri, Staiano, Sangokoya, Letouzé, & Oliver, 2016). İnsanlar karar vermek için verilere ve makine öğrenimine güvenmeye başladıkça sosyal dinamiklerin doğası da değişmeye başlamıştır.

Öncelikle veri doğası itibariyle sosyaldir. Yapay Zekanın omurgası niteliğindeki olan Makine Öğrenimi, eğitim verilerine dayalı olarak tahminler veya sınıflandırmalar yapma sürecini ifade ettiğinden makine ile sosyal çevre arasındaki etkileşimi açıklar ve oluşturur. Ancak algoritma tasarımı ve yapay zekâ uygulamasının, sonuçların nesnel ve tarafsız olup olmadığını belirleme olasılığı daha yüksektir.

Yapay zekâ tasarımı, amaçların, yani verilerin ve uygulamanın buluştuğu süreçtir. Yapay zekâ tasarımı yoluyla verilere, kalıpları ve kümeleri tespit edecek şekilde anlam verilir (Airoldi, 2021; Kaufmann, Egbert, & Leese, 2019). Tasarımda insan ve verinin iç içe geçmiş olması göz önüne alındığında, yapay zekâ pratikte yalnızca bir makine değil aynı zamanda sosyal bir aktör olarak düşünülebilir. Bu durum, en çok makine tarafından alınan kararların toplumsal etkisinde ve önyargılı sonuçlarda belirgindir (Rahwan et al., 2019). Bu bileşenlere ve bunların bütünleşik toplumsal etkilerine baktığımızda, meselenin ardındaki 'sosyolojiyi' açıkça kavrayabiliriz. Dolayısıyla bu türden bir sosyal aktörlük, sosyolojik araştırmanın meşru bir nesnesi olarak kabul edilebilir (Airoldi, 2021).

Algoritmaların ve yapay zekanın toplumsal etkisine ilişkin tartışma yeni olmamakla birlikte giderek yoğunlaşmaktadır (Joyce et al., 2021). Algoritmalar, insanların yaşamları hakkında kararlar almaları anlamında sosyal gruplar arasında 'dijital sınırlar' çizebilen güç mekanizmaları olarak düşünülebilir (Airoldi, 2021). Bu bağlamda 'Sosyoteknik sistem' terimi yapay zekanın değerlerle, kurumsal uygulamalarla ve eşitsizliklerle iç içe geçme biçimlerini ifade etmek için kullanılmaktadır (Joyce et al., 2021). Bu bakış açısı aynı zamanda yapay zekâ tasarımı ve kullanımının sosyolojik analizinin yolunu da açmaktadır.

Yapay zekâ yaşamın çeşitli alanlarına nüfuz ettikçe sosyal etkileri daha gözlemlenebilir hale gelmektedir. Örneğin, finans sektöründeki yapay zekâ uygulamaları riskleri tahmin ederek ve yatırım önererek finansal işlemleri düzenlediği için iyi bilinen bir örnektir. İstihdam süreçlerine ilişkin kararların algoritmalar tarafından alınması nedeniyle işgücü piyasası da yapay zekâ teknolojisi kullanılarak yeniden şekillenmektedir.

Sosyologların ve diğer sosyal bilimcilerin yapay zekâ konusundaki çekinceleri, yapay zekânın sosyal/toplumsal amaçlarla kullanılmasına ilişkin eleştirilere temel oluşturmaktadır. Yapay zekâ algoritmalarının verdiği önyargılı kararlar, eleştirel tartışmaların ortaya çıkmasının nedenlerinden biridir. En basit tanımıyla yanlılık, verilerdeki istatistiksel kalıpları göstermeye yardımcı olabilecek standarttan sapma olarak tanımlanabilir. (Ferrer, van Nuenen, Such, Coté, & Criado, 2020).

Algoritmik yanlılık, makine öğreniminde sürecin farklı parçalarından kaynaklanabilecek yanlılıkla ilgili her türlü sorunu ifade eden genel bir kavramdır. Yanlılığın nedenleri hem teknik hem de sosyal olabilir: Tasarımcıların ve verilerin yanlılıkları algoritmaya yansıyabileceği gibi yapay zekanın kullanımı, halihazırda var olan eşitsizlikleri ve yanlılıkları yeniden üretebilir (Livingston, 2020). Bu bağlamda yanlılığın temel nedenlerini insan kaynaklı yanlılık ve veri kaynaklı yanlılık olmak üzere iki kategori altında tartışmak mümkün ve gereklidir.

Yapay zekâ tasarımı ve veri işleme sürecinde insan müdahalelerinden kaynaklanan yanlılık, insanla ilgili yanlılık olarak tanımlanabilir. Bu yanlılığın bir biçimi spesifik yapay zekanın tasarımıyla ilgilidir. Tasarımcılar çeşitli ve ham haldeki veriye anlam atfederek makine öğrenimine katkıda bulunur. Yapay zekâ geliştiricileri, posta kodlarının sosyal statü göstergesi olması gibi belirli verilerin gerçek sosyal anlamının farkında olmayabilir (Joyce et al., 2021). Bu gibi durumlar insandan kaynaklı yanlılığa sebep olabilmektedir.

Makina öğrenimi ve yapay zekada kullanılan eğitim verileri ile gerçek zamanlı veriler farklı sorunları içerebilir. Makina öğreniminde kullanılan eğitim verileri çoğu zaman önyargılarla yüklüdür ve bunlar yapay zekaya taşınır (Ferrer et al., 2020). Veriler, işe alımda cinsiyet ayrımcılığı gibi 'yaygın, kalıcı toplumsal düzeydeki önyargıları’ yansıtabilir (Fountain, 2022). Öte yandan, veriyi toplayan uygulamalara ve kurumlara olan güvenin azalması gibi çeşitli nedenlerden dolayı kişilerin veri paylaşma eğilimi eşit şekilde dağılmamaktadır. (Kontokosta & Hong, 2020). Bireyler, veri işlemenin gizliliği veya şeffaflığı endişesiyle belirli uygulamaları kullanmaktan veya her türlü kişisel veriyi paylaşmaktan kaçınabilirler.

'Sosyal Fayda için Yapay Zekâ' (AI4SG), temel olarak ticari olmayan ve toplumsal açıdan faydalı sonuçlar üretmeyi amaçlayan yapay zekâ projelerine atıfta bulunmak için kullanılan bir terimdir (Cowls, Tsamados, Taddeo, & Floridi, 2021). Yapay zekâ alanının sosyal sorunlara yönelmesi, temel olarak sosyal hizmet, eğitim, sağlık ve diğer birçok sektördeki verileştirmenin, sosyal sorunların daha iyi anlaşılmasını ve etkili çözümler bulunmasını kolaylaştırabileceği varsayımıyla ilgilidir. (Coulton, Goerge, Putnam-Hornstein, & De Haan, 2015).

Yapay zekâ ile sosyal faydayı ölçülebilir bir şekilde bir araya getirmek için çeşitli ilkeler önerilmiştir. Örneğin, kapsamı basitleştirmek amacıyla AI4SG projeleri için önkoşul olarak üç temel gereklikten bahsedilebilir; (1) insan yaşamına ve/veya doğaya zarar veren sorunları önlemek, azaltmak ve/veya çözmek veya (2) sosyal açıdan arzu edilen veya çevresel açıdan sürdürülebilir olan gelişmeleri kolaylaştırmak ve (3) yeni zararlara yol açmamak ve/veya mevcut eşitsizlikleri genişletmemek (Cowls et al., 2021).

Çağımızın birçok sosyal, ekonomik ve çevresel sorununu çözmeye yönelik bir yaklaşım olan AI4SG, çeşitli bakış açılarından eleştirilmektedir. Kişisel bilgilerin verileştirilmesi ve verilerin insani amaçlarla kullanılması, mülteciler veya afet mağdurları gibi halihazırda dezavantajlı gruplar söz konusu olduğunda kırılganlıkları artırabilir (Coulton et al., 2015; Magalhães & Couldry, 2021). Belirli bir sorun nüfusun küçük bir bölümünü etkiliyorsa veya etkilenen grup veri setinde yeterince temsil edilmiyorsa söz konusu sorun fark edilmeyebilir (Coulton et al., 2015). AI4SG projeleri aynı zamanda dikkatlerin sosyal ve çevresel sorunların temel nedenlerinden uzaklaşmasına neden oldukları için de eleştirilmektedir. Teknoloji şirketlerinin (çözüm üretiminden çok) yapay zekâ yeteneklerini ön plana çıkardıkları iddiası da bir diğer eleştiri konusudur (Holzmeyer, 2021).

Yapay zekâ uygulamalarında ve algoritmalarda yanlılığın azaltılması ya da engellenmesine ilişkin çeşitli öneriler bulunmaktadır. Teknik açıdan bakılırsa daha iyi algoritmalar oluşturmaktan bahsetmek mümkün olsa da böyle bir çözüm bile sorunun doğasına, yani teknolojinin sosyal/sosyolojik özelliğine dair yeni bir bakış açısı gerektirmektedir. Şeffaflık ise, veri toplama ve tasarım aşamaları da dahil olmak üzere birçok pratik önkoşulu içeren kavramsal bir çözüm olarak önerilmektedir (Boscoe, 2019). Güçlü insan denetimi standartları da yapay zekâda veriyle ve hatta insanla ilgili yanlılığı en aza indirecek bir süreç olarak düşünülmektedir (Taddeo & Floridi, 2018). Kapsayıcı yapay zekâ ise, yapay zekâdaki önyargının (ve ayrımcılığın) üstesinden gelmek için çözüm olarak önerilen bir diğer kavram/ilkedir.

Yanlılık sorununun üstesinden gelmeye yönelik teknik önlemler ve sistemin iyileştirilmesi ancak yapay zekânın insani yönünün güçlendirilmesiyle anlamlı olabilir. Tüm teknolojik yenilikler gibi yapay zekânın da insani ve toplumsal faydaya hizmet etmesi gerektiği unutulmamalıdır. Yukarıdaki tartışmaların ışığında gelecekteki araştırmalar, yapay zekâ ve algoritma yanlılığının insanla ilgili temel nedenlerinin ayrıntılı analizine ve evrensel yönergeler ile kapsayıcı değerlerin somutlaştırılmasına odaklanmalıdır. Bu da tüm paydaşların ve yararlanıcıların dahil olduğu kapsamlı bir tartışmayı gerektirir.