İstanbul Üniversitesi Sosyoloji Dergisi 44(1): 221-238

DOI: https://doi.org/10.26650/SJ.2024.44.1.0001
http://iusd.istanbul.edu.tr

İstanbul Üniversitesi Sosyoloji Dergisi

Submitted: 5 January 2024
Accepted: 25 March 2024
Online Published: 3 May 2024

RESEARCH ARTICLE

Participatory Management Can Help Artificial Intelligence Ethics Adhere to Social Consensus

Mahmut Özer [1], **Matjaz Perc** [2,3,4,5,6], **Hayri Eren Suna** [7]

Abstract
Artificial Intelligence (AI) is rapidly becoming pervasive, reshaping social structures, cultural dynamics, and labor markets. This rapid growth has ignited global discussions surrounding AI's challenges, including its tendency to perpetuate biases and social inequalities, ignoring societal values, and affect diverse sectors such as genetics, drug production, defense, and democratic processes. This study examines AI ethics within the framework of social consensus, advocating for participatory management as a crucial approach to address these challenges. The proposed methodology includes the entire AI lifecycle, promoting inclusive practices from initial design through implementation, monitoring, and control. The participatory management model is structured in three phases: Stakeholder Engagement, which advocates for the active involvement of diverse stakeholders in the development of AI systems to ensure a range of perspectives in design, modeling, and implementation; Monitoring and Alignment, which emphasizes the continual observation of AI systems' interaction with their environments; and Macro Level Impact Analysis, which evaluates the broader societal impacts of the AI ecosystem across domains such as education, culture, health, and safety. This study underscores the importance of a collaborative, inclusive approach in AI development and management, emphasizing the need to align AI advancements with ethical principles and societal well-being.
Keywords: Artificial intelligence • ethics • bias • fairness • participatory algorithmic management • algorithmic accountability

1 Mahmut Özer (Prof. Dr.), National Education, Culture, Youth and Sports Commission, Ankara, Türkiye.
E-mail: mahmutozer2002@yahoo.com ORCID: 0000-0001-8722-8670
2 Matjaz Perc (Prof. Dr.), Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia.
E-mail: matjaz.perc@gmail.com ORCID: 0000-0002-3087-541X
3 Complexity Science Hub Vienna, 1080 Vienna, Austria
4 Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 404, Taiwan
5 Alma Mater Europaea, Slovenska ulica 17, 2000 Maribor, Slovenia
6 Department of Physics, Kyung Hee University, 26 Kyungheedae-ro, Dongdaemun-gu, Seoul, Republic of Korea
7 **Correspondence to:** Hayri Eren Suna (Dr.), Ministry of Education, Paris Embassy Educational Office, Paris, France.
E-mail: herensuna@gmail.com ORCID: 0000-0002-6874-7472
To cite this article: Özer, M., Perc, M., & Suna, H. E. (2024). Participatory management can help artificial intelligence ethics adhere to social consensus. *İstanbul Üniversitesi Sosyoloji Dergisi, 44,* 221-238. https://doi.org/10.26650/SJ.2024.44.1.0001

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License

For a long time now, intensive data production has been taking place in all areas of life, with an accompanying rise in data collected and analyzed. Humanity's inclination toward employing scientific methodology to comprehend and control phenomena has steadily grown, necessitating the creation of reliable and quantitative indicators (Ye, 2017). Advancements in scientific research have led to an unprecedented proliferation of information and data. The total volume of data generated, which stood at 2 zettabytes in 2010, has increased exponentially to 120 zettabytes in 2023, with projections estimating it to reach 181 zettabytes by 2025 (Digital Center, 2023). Moreover, there is clear evidence of exponential growth in scientific research output and the accumulation of scientific data in recent years (Bornmann et al., 2021; Pelacho et al., 2020).

The massive scale of generated data has facilitated the development of intelligent systems, particularly with the involvement of artificial intelligence (AI), machine learning, and deep learning, thus significantly accelerating this process (Perc et al., 2019). These technologies have propelled machines beyond simply being designed to perform routine tasks following the purposes for which they were designed (Daugherty and Wilson, 2018; Manyika and Sneader, 2018). Because of these technologies, machines that can adapt to new situations can be created, as well as machines that can be designed to "learn to learn" and adapt to new methods of achieving their goals. By employing the data and experience they have gained over time, machines can, for example, be empowered to perform their tasks more effectively through machine learning methods. Through the use of the new data gained from machine learning methods, it is possible to further minimize human intervention in machines (Soori et al., 2023). Deep learning methods allow machines to acquire knowledge through neural networks by mimicking the learning process of the human brain (Shinde and Shah, 2018). As the name implies, AI includes all the methodologies employed to acquire, generate, and refine innovative tools using various sub-methods such as machine learning and deep learning (Soori et al., 2023).

AI and automation technologies have become pervasive across various aspects of life, spanning from the economy to healthcare, security to education, with their application continually expanding. AI systems are constantly evolving, interconnecting with one another to form a new state known as an AI ecosystem comprising socio-technical systems (Stahl, 2023). Alongside studies on how the AI ecosystem affects and reshapes societies, there is now widespread discourse on the potential risks it may pose (Crawford and Calo, 2016; Harari, 2017; Suleyman, 2023). The dual nature of opportunities and threats has created a multifaceted perception of AI across diverse segments of society (Bozkurt and Gursoy, 2023; Brauner et al., 2023; Gerlich, 2023).

The emergence of the AI ecosystem has led to transformative shifts in the labor market, surpassing previous technological revolutions and fundamentally altering labor dynamics (Acemoğlu and Restrepo, 2018; Frank et al., 2019). Consequently, AI systems have brought about transformations across nearly all professions within the labor markets, completely reshaping expectations for skills and occupations (Harari, 2017; Özer, 2024). Numerous studies aim to determine the extent of this transformation in professions, yielding divergent findings that span from pessimistic assessments predicting the disappearance of many professions in the labor market (Arntz et al., 2016; Pajarinen et al., 2015) to optimistic appraisals suggesting that the labor market transformation in the emergence of new professions (Aghion and Howitt, 1990; 1994; Bartelsman et al., 2004). However, Frank et al. (2019) argue that both assessments are incomplete, attributing this incompleteness to factors such as a lack of high-quality data on the nature of work, deficiencies in experimental models concerning micro-level processes, and inadequate understanding of how cognitive technologies interplay with economic dynamics and institutional mechanisms.

Aside from these discussions, it is acknowledged that the impact of AI systems on labor markets is profoundly impactful, making the next significant milestone following previous technological advancements such as the widespread use of clocks, coupled with the establishment of a shared understanding of time, have played an important role in shaping the contemporary structure of the labor market (Thompson, 1967). Global standardization of time facilitated meticulous production planning and enabled distribution arrangements on a global scale. Similarly, mechanization has led to the demise of labor-intensive production methods and brought about significant changes in labor market dynamics (Montesano, 2011; Samuelson, 1988). With AI applications increasingly permeating various sectors, it is poised to catalyze one of the most profound changes in the labor market since these advancements.

As AI and data-driven systems continue to proliferate, we have gained extensive insight into their operations. However, the convergence of AI systems with automation, leading to their dominance and potential takeover of the labor market, presents a new dilemma: the replacement of human labor with machine-driven employment. The prevalent inclination in labor markets toward automation over human-complementary approaches poses a risk of shifting the balance between humans and machines, potentially increasing unemployment, inequality, and disrupting social harmony (Acemoğlu et al., 2023). Therefore, developing new policies that ensure the integration of these systems into labor markets in a manner that strengthens the human-complementary approach, rather than solely focusing on the economic advantages of AI systems, can mitigate their adverse long-term negative impacts on societal structures (Capraro et al., 2023).

Simultaneously, there is a growing discourse surrounding how these systems, while offering significant advantages, also contribute to heightened inequalities, biased behaviors, or outcomes that contradict societal ethical values (Rahwan, 2018; O'Neil, 2016; Suleyman, 2023). Inequalities often stem from disparate access to data, as certain groups, institutions, and organizations with access to comprehensive datasets also wield disproportionate influence over data analysis and control (Boyd and Crawford, 2012).

However, the ethical debates surrounding the use of autonomous vehicles and AI systems in military, genetic, biotechnology, and health fields highlight the potential costs of AI systems that ignore societal values while operating autonomously (Citron and Pasquale, 2014; Rahwan, 2018; 2019). Intense discussions are taking place on the necessity for AI and data-driven systems to consider societal values, as their neglect could lead to social conflicts, such as perpetuating biases (Lee et al., 2019; Rahwan, 2018; Stahl, 2023). Thus, a profound discussion ensues on how and to what extent societal values can be incorporated into the design of decision-making mechanisms in AI systems (Piano, 2020).

Since the emergence of the first AI applications, sociological evaluations of their impact on the social context have proliferated. As machines increasingly exhibit "intelligence," crucial discussions regarding their social consequences (Schwarz, 1989). Berman (1989, 1992) suggests that AI's rapid development holds the potential to be developed within a short period of time and that its development will allow for a greater control over society by elites (Liu, 2020). Turkle (1984) previously observed that the blurring of distinctions between humans and machines would occur as AI models increasingly mimic human learning processes, thus fostering a propensity toward analytical thinking and the utilization of machine learning techniques Woolgar (1985) emphasizes the "social" attributes that distinguish humans from machines, advocating for evaluations that transcend conventional sociological approaches. Decades ago, the eminent sociologist Tarde foresaw the utilization of behavior statistics for predictive purposes (Didier, 2015). In contrast to Heidegger and his predecessors, Foucault eschewed an explicit distinction between machine and mind, instead introducing the concept of the "intelligent machine" (Hernandez-Ramirez, 2017). Luhmann, another influential sociologist, acknowledged the convergence between humans and machines, positing that as long as these interactions yield meaningful information, societies will continue to exist by undergoing various transformations (Wolfe, 1991). The proliferation of AI applications has notably enriched sociological assessments. Floridi (2011) contends that AI applications have fundamentally changed individuals' self-assessment, interpersonal communication, and interaction with the outside world. Presently, machine learning results intersect with numerous social factors, including the repetition of cultural and social stereotypes (Arseniev-Koehler and Foster, 2022; Boutyline et

al., 2023), increased political opposition (Steward et al., 2020), and increased inequalities (Joyce et al., 2021).

Therefore, this study delves into conceptual approaches aimed at addressing the ethical concerns surrounding AI systems, subjecting them to thorough discussion and evaluation. In addition, the participatory management approach, attuned to the advocated values in the development of AI systems, is being broadly examined.

Ethical Issues Arising from AI Systems

One of the foremost concerns surrounding AI systems revolves around data protection and privacy. As the AI ecosystem expands, the realm of privacy is diminishing, with an increased risk of identifying supposedly anonymous personal data (Zimmer, 2008). Individuals may unknowingly have their data used (Lewis et al., 2008), leading to ethical debates regarding whether the use of any available data is ethical (Boyd and Crawford, 2012).

However, these systems often introduce bias into their generated results, causing them to deviate from fairness. As a result of its biased results and predictions in favor of male applicants, Amazon, for example, stopped using AI-based hiring software (Mu, 2023). It has been determined that the DALL-E mini platform, which is an AI-based platform that can draw pictures based on commands, continuously creates male profiles in response to commands such as "manager" and "CEO" (Wan and Chang, 2022). Moreover, a digital management system frequently used in the healthcare industry in America predicted that black individuals are more likely to miss appointments or arrive late (Sjoding et al., 2021). All of these examples demonstrate that algorithmic decisions that lead to important decisions in different fields are particularly biased.

Biases within AI systems can originate not only from assumptions and indicators relevant to the field but also directly from the datasets used for learning *(training datasets)*. AI systems are developed based on human-made assumptions, defined limits, and features, which can initially introduce bias into the system design (Erdi, 2020; Piano, 2020). Moreover, biases are often reinforced by the training datasets.

Societal infiltration permeates these systems through the datasets utilized, leading to the manifestation of biases in the behaviors of developed intelligent systems, thereby perpetuating inequalities. Consequently, the efficacy of an algorithm is contingent upon the quality of the data it uses (Barocas and Selbst, 2016). Furthermore, the results generated by algorithms can be affected by factors unrelated to the dataset, such as estimation methods. Consequently, if the behaviors exhibited by AI algorithms remain uncorrected, biases inherent in the training dataset can be transferred to the AI algorithm, thus reflecting in its decisions (Aquino, 2023; Ntoutsi et al., 2020; Özer et al., 2024).

This underscores the risk of reproducing and reinforcing societal biases within AI and data-driven systems (Lum and Isaac, 2016).

Therefore, the central focus of debate within the scope of AI ethics revolves around the fairness of AI systems' behaviors. The tendency of AI systems to demonstrate behaviors that favor certain groups over others has prompted discussions regarding their impartiality and underlying origins. Extensive research has been conducted on this issue, revealing instances of bias that result in discrimination outcomes against black individuals. Even when black offenders pose no risk of reoffending, they are erroneously classified as having a higher likelihood of reoffending compared with white individuals, with a misclassification rate of approximately double (Angwin et al., 2016).

Similarly, software incorporating AI algorithms intended to pinpoint potential future crimes and offenders has demonstrated biases inherent in the datasets. In a particular state, despite the presence of the issue across all areas, police patrols predominantly targeted regions inhabited by nonwhite and socioeconomically disadvantaged individuals, influenced by biases within the datasets used by these algorithms (Lum and Isaac, 2016). In this case, the bias reflects real-world disparities that are transferred to the algorithms through the learning data on which they rely. Consequently, the concentration of policing efforts in these areas increases the likelihood of individuals from these communities being apprehended. Consequently, crime data in these areas increases, prompting the algorithm, which continuously learns from new data, to redirect police patrols to the same area. Thus, inequalities are continuously reinforced through a feedback loop (Lum and Isaac, 2016). Therefore, when the dataset used to update the algorithm is biased, it creates a vicious cycle that perpetuates inequalities (O'Neil, 2016).

Similar biases are also manifesting in the field of healthcare, where algorithms are observed to prioritize care for individuals with greater access to healthcare services (Bates et al., 2014; Obermeyer et al., 2019). In such instances, the algorithm may reduce follow-up screenings for those with limited access to healthcare services, thereby increasing health risks for disadvantaged populations (Mittermaier et al., 2023). For instance, Obermeyer et al. (2019) conducted a study revealing that algorithms commonly use healthcare expenditure data to determine access to healthcare. In this context, white individuals disproportionately benefited more from high-risk care programs than other racial groups. Correcting this bias led to a substantial increase in the additional services that black patients would receive, from 17.7% to 46.5%.

In summary, biases within AI algorithms transcend beyond mere technical errors; they serve as a direct reflection of societal power structures, economic disparities, and political systems (Ulnicane and Aden, 2023). Thus, the Matthew effect warrants discussion

and is often cited in the context of the reproduction of societal inequalities. Numerous studies have established the application of the Matthew effect to social phenomena (Strevens, 2006), embodying the adage from the Gospel of Matthew, ”More will be given to those who have.” Within this framework, individuals in society occupy diverse positions based on their resource allocation, with these discrepancies perpetuating advantages and disadvantages (da Silva, 2021; Perc, 2014). The Matthew effect has been shown to accentuate advantages for privileged segments of society while exacerbating disadvantages for disadvantaged segments (Merton, 1968; Zuckerman, 1989; Özer, 2023a; Özer, 2023b; Özer, 2024; Özer and Perc, 2020; Özer and Perc, 2021; Rigney, 2010). Given their ability to reflect and expand social inequalities, AI algorithms can enhance the Matthew effect. Consequently, inequalities based on gender, race, socioeconomic status, etc., are perpetuated within society (O’Neil, 2016).

Participatory Management of AI Systems

AI systems have been the subject of numerous ethical debates, with various observations and solution proposals put forth. To consolidate these disparate solutions and establish a comprehensive framework, Stahl (2023) defines the concept of “responsible AI”, conceiving the structure as an AI ecosystem consisting of socio-technical systems. Stahl (2023) situates all efforts within the scope of system responsibilities, encapsulating them under the concept of ”meta-responsibility in the ecosystem.” Therefore, it becomes necessary to conduct all stages of each application within the AI ecosystem with a heightened awareness of societal implications, from the design phase to implementation. Addressing both the social impact of systems and society’s influence on the system necessitates adopting a social-systems approach (Crawford and Calo, 2016).

In this context, the primary focus is often directed toward algorithms, which are frequently considered black boxes,” with recommendations advocating for their openness and transparency to all stakeholders (Mayer-Schönberger and Cukier, 2014; Pasquale, 2015). Some even advocate the use of open-source software in AI and machine learning applications (Thimbleby, 2003). However, concerns have been raised regarding the potential adverse effects of code transparency, such as reducing efficiency for code programmers, negative impacts on competition, and exposure to sensitive data (de Laat, 2018; Piano, 2020; Sonnenburg, 2007). Open access to data can result in the depreciation of its commercial value, given the considerable effort invested in curating datasets for commercial analysis and usage. Consequently, institutions developing datasets for commercial purposes and employing them in AI applications may incur financial losses. Furthermore, open access may inadvertently expose certain copyrighted features or require permissions, especially when datasets are compiled from multiple sources of data. There are arguments that complete transparency could

raise issues concerning trade secrets or render the system vulnerable to gaming and manipulations (Pasquale, 2011; Diakopoulos, 2015; Rahwan, 2019).

Rahwan (2018) argues that the source code of algorithms offers limited insights into the behaviors of these applications. Therefore, he suggests that instead of focusing on the code, a more effective approach is to continuously monitor the behaviors of algorithms, i.e., the results they yield. In healthcare, in addition to adaptation studies, there is a notable emphasis on ongoing monitoring of AI application behavior (Kostick-Quenet and Gerke, 2022). Similarly, in journalism, there is a growing trend of contributing to algorithmic accountability by using a reverse engineering approach, starting from the algorithm's output (Diakopoulos, 2015). By developing methodologies that analyze algorithmic outcomes to determine their underlying mechanisms and operational methodologies, it is possible to develop more accountable and transparent algorithms. Hence, observing machine behavior within AI systems is of paramount importance alongside algorithms and data (Rahwan et al., 2019). Continuous monitoring of the behavior of AI systems facilitates the identification of biases and deviations arising from both design choices and the datasets used in learning, enabling corrective measures to be implemented. When datasets are tailored to very specific groups, diminishing their representational power, decisions may exhibit bias against the unrepresented, underrepresented, or overrepresented. Therefore, it remains an ongoing imperative to scrutinize whether algorithms yield undesirable outcomes for particular societal groups, either due to underlying assumptions or the training dataset (Nazer et al., 2023). The most effective approach to mitigating the effects of representational and measurement biases involves using key variables that are inherently unbiased and collecting datasets of higher quality (Baker and Hawn, 2021).

Etzioni and Etzioni (2016) advocate the development and implementation of second-order AI programs known as "ethics bots" within this framework, tasking with guiding AI algorithms. This approach empowers individuals to steer the intelligent systems they use according to their values and preferences. Furthermore, these developed bots can extend their utility to other intelligent applications employed by individuals. However, as ethics bots account for moral preferences manifested in real behavior (Etzioni and Etzioni, 2016), a significant challenge arises due to the weak correlation between attitudes and actual behaviors (Azjen et al., 2004). Consequently, the guidance provided by ethics bots, though rooted in individuals' actual behaviors, may contradict their judgments. Conversely, such approaches are likely to instigate discussions concerning data security and privacy (Boyd and Crawford, 2012).

The ability to intervene in behaviors that perpetuate discrimination based on factors such as gender and socioeconomic status presents an opportunity to rectify the conduct of AI systems. The increasing focus on bias or discrimination in the behaviors of widely used algorithms prompts a more comprehensive examination of the issue. These

discussions, in essence, advocate the adoption of participatory algorithmic management as a more fundamental solution for improving algorithms.

Here, a detailed introduction to the participatory algorithm model is necessary. Participatory algorithm models are characterized by the participation of several stakeholders in the design process, including users, across all phases of the development process (Birhane et al., 2022; Bratteig and Verne, 2018; Gerdes, 2022; Hussain et al., 2012; Hossain and Ahmed, 2021). As a result of this approach, AI algorithms undergo evaluation from a large number of different perspectives, with diverse social dynamics considered during the development phase (Delgado et al., 2023; Hossain and Ahmed, 2021). Several social values, perspectives, and qualities can manifest in algorithms when representatives from different segments of society engage in the design development process. Ensuring broad participation in algorithm development is crucial for mitigating inequalities. Several studies have illustrated that algorithms lacking adequate representation often yield biased results across various areas (Akter et al., 2021; Fu et al., 2022; Hussain et al., 2012; Panch et al., 2019). Moreover, such biases disproportionately affect marginalized and underrepresented groups (Hossain and Ahmed, 2021).

Birhane et al. (2022) emphasize the importance of using a participatory algorithm model when making high-stake decisions. Usually, in scenarios where data-driven decisions do not carry a significant amount of risk, biased inferences can aid in understanding the algorithm's biases and refining them to create more unbiased results (Gerdes, 2022). However, the ramifications become perilous, especially when these results inform critical decisions concerning individuals, and the adverse effects of algorithmic flaws are substantial (Birhane et al., 2022; Gerdes, 2022). Additionally, Bondi et al. (2021) emphasize that the participatory model should form an integral part of the social dimension of AI. According to them, algorithms must undergo evaluation by several stakeholders to ensure that results are equitable and unbiased.

Furthermore, Bratteig and Verne (2018) argue that the establishment of a participatory algorithm is paramount for fostering more egalitarian outcomes. However, they highlight two key challenges that must be addressed. First, there is the difficulty of elucidating the technical aspects of AI technologies and the overarching framework of AI to stakeholders involved in the model's development (Bratteig and Verne, 2018). Second, predicting the ramifications of alterations to algorithm structures proves challenging, given algorithms' continual learning and generation of new information. These challenges indicate the necessity of integrating a participatory model into algorithm development from its inception and ensuring that stakeholders are informed about the subject as the development processes.

In participatory modeling, researchers, technology users, and groups affected by the modeling—essentially those involved within the modeling context—share power

and control over the design (Muller, 2009). For example, Lee et al. (2019) illustrated stakeholder involvement in an algorithmic model, establishing the "WeBuildAI" framework. An essential aspect of this framework is stakeholder involvement at all stages of model development. Initially, stakeholders are surveyed to determine the essential relevant features/variables to incorporate into the algorithm (Lee et al., 2019). Upon consensus on these features/variables, the second stage involves selecting methods for model development and decision-making. In the final stage, all participants are provided with clear indicators explaining their contribution to the entire process (Lee et al., 2019). By involving diverse stakeholders, significant biases can be reduced throughout all processes, from variable selection to model operation. The model was applied to a real-world scenario involving donation distribution, yielding a participatory algorithm that balanced distance efficiency and fair distribution constraints. Results showed that the algorithm facilitated a fairer and efficient distribution without increasing transportation distance, effectively reaching segments with higher poverty rates, lower incomes, and limited access to food.

In this context, the Writers Guild of America emphasizes the need to consider the use of previous writings in the training of AI algorithms within the context of copyright. They assert that failure to do so leaves labor unprotected (Calacci, 2023). From a copyright perspective, writings included in AI training data are individual works, and using them without adhering to standard copyright procedures could potentially constitute a violation. Essentially, previous labor becomes raw material for AI algorithms, which require continual ingestion of new datasets to enhance their efficiency. Unions advocating for the recognition of this labor as copyright material view negotiations regarding AI algorithm features as a critical issue. Otherwise, they oppose the use of their productions as raw materials for improving AI algorithms.

Addressing this issue in AI and data-driven systems necessitates an algorithmic social consensus involving numerous stakeholders mediated by these systems. In this context, Rahwan (2018) proposed establishing a "society-in-the-loop" framework that integrates societal values for general applications affecting larger social segments. This approach mirrors the narrow "human-in-the-loop" approach used to rectify errors and optimize AI systems in more confined applications. Within this framework, methodologies such as "value-sensitive design," previously applied in various system designs, have been integrated into AI systems (Aldewereld, 2014; Friedman, 1996; Rahwan, 2018). Another proposed approach for embedding social values in algorithms is crowdsourcing (Bonnefon et al., 2016; Conitzer et al., 2015; Liu, 2012). Termed "crowdsourcing," this process involves gathering opinions from stakeholders deemed significant within a relatively short period of time and implementing improvements based on this feedback (Bonnefon et al., 2016). By collecting feedback from various segments of society regarding AI application outcomes, this method proves valuable

in reducing algorithm biases. This facilitates the evaluation of opportunities for measuring social values for integration into algorithms.

One proposed solution for ensuring that AI systems align with social consensus involves employing professional algorithm auditors. This approach also allows the interaction between human and AI, and places the human control as a support and control mechanism for AI outputs. However, Rahwan (2018) cautions against the possibility of AI systems bypassing such audits, emphasizing the need for caution. In fact, initial studies in journalism indicate that computational journalists possessing technical skills can serve as algorithm auditors representing society's interests (Diakopoulos, 2015).

While research in this area is still in the nascent stages, participatory management not only ensures algorithmic effectiveness but also fosters moral integrity. Thus, it provides an opportunity to prevent the perpetuation of social inequalities arising from algorithm biases, thereby averting outcomes detrimental to various groups or stakeholders (Friedman and Nissenbaum, 1996; Lee et al., 2019; Zarsky, 2016).

On the other hand, AI systems operate within intricate networks, interacting with one another and with humans in real-life situations. Emergent behaviors resulting from these interactions between intelligent systems and humans serve as primary determinants. Consequently, assessing how these systems affect and transform behaviors within a hybrid framework including humans, machines, and the entire ecosystem is of great importance (Rahwan et al., 2019). Recent discussions have emerged regarding how intelligent machines, through mutual interactions within this ecosystem, influence cultural evolutionary processes and how, in collaboration with humans and machines, this transformation can foster the development of a shared and harmonious culture (Brinkmann et al., 2023).

AI systems can generate results based on their own experiences, with changes in behavior stemming from these experiences commonly observed in financial and commercial applications, especially in recommendation algorithms and AI attacks (Biggio, 2013; Newmyvaka et al., 2006; Parkes and Wellman, 2015; Rahwan et al., 2019; Tramer et al., 2017). Thus, the continual interaction among AI systems and between these systems and humans not only influences their behaviors but also engenders collective behaviors based on these interactions (Rahwan et al., 2019). Consequently, beyond merely controlling individual behaviors discretely, dynamic oversights of this ecosystem comprising AI systems and humans prove indispensable.

Conclusion

The AI ecosystem's transformative impact on society, functioning as a socio-technical system, is profound and reciprocal. AI systems not only influence human

behavior but are also shaped by it, leading to a dynamic, interdependent ecosystem. This interaction underscores the necessity of a holistic approach to understanding and managing AI, where ethical considerations are integral not only in the design and implementation of AI components but also in ongoing monitoring and evaluation.

A key strategy to ensure the ethical operation of the AI ecosystem and mitigate bias is the adoption of a participatory management approach. This model, derived from various solutions proposed in the literature, advocates for widespread adoption. In this study, we outline a three-stage participatory management model for the AI ecosystem.

1. Stakeholder Engagement: In the initial stage, the active involvement of stakeholders is paramount, including their participation in designing, modeling, selecting learning data, and testing AI systems. This comprehensive engagement not only amplifies the diversity of perspectives but also facilitates the early detection of biases. At each stage of algorithm development and implementation, each stakeholder should be allowed to voice their insights. It is essential to offer participants a platform to express their opinions on project progress and outcomes. This inclusive approach includes a diverse array of societal groups, ensuring that diverse perspectives are considered. For example, engaging healthcare professionals in the development of medical AI can ensure the integration of patient-centric values. The objective of this stage is to minimize biases, enable value-sensitive design, and proactively address potential societal implications.

2. Continuous Monitoring and Adaptation: AI systems evolve dynamically in response to their environment and interactions. Fundamentally, AI algorithms are in a perpetual state of learning and transformation, incorporating each outcome and new information as additional data points. Consequently, the potential for biased outcomes always exists, underscoring the necessity for continuous monitoring. This includes algorithmic auditing, which is crucial for identifying and rectifying potential errors or undesirable behaviors in AI systems after deployment. For instance, regular audits of an AI recruitment tool ensure its ongoing adherence to nondiscriminatory practices over time.

3. Macro Level Ecosystem Analysis: In addition to monitoring individual AI systems, developers and relevant stakeholders must conduct a comprehensive analysis of the overarching AI ecosystem within a transparency framework. This holistic approach enables the assessment of the ecosystem's influence on societal structures, culture, health, safety, etc., at a macro level. Achieving this requires collaboration among software engineers, humanities scholars, and behavioral scientists (Awad et al., 2020). Analogous to evaluating the collective impact of social media AI algorithms on public discourse, this task is challenging yet imperative for understanding the broader societal implications of AI deployment.

In conclusion, the proposed participatory management model offers a comprehensive framework for ethically guiding the development and implementation of AI systems. By incorporating diverse stakeholder input, ensuring continuous monitoring, and evaluating the broader ecosystem impact, AI advancements can be better aligned with societal values and ethical standards.

Peer-review: Externally peer-reviewed.
Conflict of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Grant Support: The authors received no financial support for the research, authorship, and/or publication of this article.
Author Contributions: Conception/Design of Study: M.Ö., M.P., H.E.S.; Data Acquisition: M.Ö., M.P., H.E.S.; Data Analysis/Interpretation: M.Ö., M.P., H.E.S.; Drafting Manuscript: M.Ö., M.P., H.E.S.; Critical Revision of Manuscript: M.Ö., M.P., H.E.S.; Final Approval and Accountability: M.Ö., M.P., H.E.S.

References

Acemoğlu, D., & Restrepo, P. (2018). *Artificial intelligence, automation and work*. NBER Working Paper 24196. National Bureau of Economic Research.

Acemoğlu, D., Autor, D., & Johnson, S. (2023). Can we have pro-worker- AI? Choosing a path of machines in service of minds. *CEPR Policy Insight,* No.123, 1-12.

Aghion, P., & Howitt, P. (1990). *A model of g*rowth through c*reative destruction.* NBER Working Paper 3223 National Bureau of Economic Research.

Aghion, P., & Howitt, P. (1994). Growth and unemployment. *Rev Econ Stud*, 61, 477–494.

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management,* 60, 102387.

Aldewereld, H., Dignum, V., & Hua Tan, Y. (2014). Design for values in software development. In Jeroen van den Hoven, Pieter E. Vermaas, I. v. d. P., (Eds), *Handbook of ethics, values, and technological design*. Springer.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (23 May, 2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica.

Aquino, Y. S. J. (2023). Making decisions: Bias in artificial intelligence and data-driven diagnostic tools. *Australian Journal of General Practice*, 52(7), 439-442.

Arntz, M., Gregory, T., & Zierahn, U. (2016). *The risk of automation for jobs in OECD countries: A comparative analysis*. OECD Social, Employment and Migration Working Paper 189.

Arseniev-Koehler, A., & Foster, J. G. (2022). Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. *Sociological Methods & Research*, 51(4), 1484-1539.

Awad, E., Dsouza, S., Bonnefon, J. F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing: Moral machines. *Communications of ACM*, 63(3).

Azjen, I., Brown, T. C., & Carvajal, F. (2004). Explaining the discrepancy between intentions and actions: The case of hypothetical bias in contingent valuation. *Personality and Social Psychology Bulletin,* 30(9), 1108-1121.

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32, 1052-1092.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review,* 104, 671-732.

Bartelsman, E., Haltiwanger, J., & Scarpetta, S. (2004). *Microeconomic evidence of creative destruction in industrial and developing countries.* The World Bank.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs,* 33(7), 1123-1131.

Berman, B. (1989).The computer metaphor: Bureaucratizing the mind. *Science as Culture,* 1(7), 7–42.

Berman, B. (1992). Artificial intelligence and the ideology of capitalist reconstruction. *AI & Society*, 6(2), 103–114.

Biggio, B. et al. (2013). Evasion attacks against machine learning at test time. In *Proc Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 387-402).

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed. S. (2022). Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (EAAMO '22). Association for Computing Machinery, New York, NY, USA, Article 6, 1–8.

Bondi, E., Xu, L., Acosta-Navas, D., & Killian., J. A. (2021). Envisioning communities: A participatory approach towards AI for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '21). Association for Computing Machinery, New York, NY, USA, 425–436.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576.

Bornmann, L., Haunschild, R., & Mutz, R. (2021).Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications,* 8, 224.

Boutyline, A., Arseniev-Koehler, A., & Cornell, D. J. (2023). School, studying, and smarts: Gender stereotypes and education across 80 years of American print media, 1930–2009. *Social Forces,* 102(1), 263–286.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society,* 15(5), 662–679.

Bratteteig, T., & Verne, G. (2018). Does AI make PD obsolete? Exploring challenges from artificial ıntelligence to participatory design. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial,* 2, 8, 1–5.

Brauner, P., Hick, A., Philipsen, R., & Ziefle, M. (2023). What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI. *Frontiers in Computer Science*, 5, 1113903.

Bozkurt, V., & Gürsoy, D. (2023). The artificial intelligence paradox: Opportunity or threat for humanity?. *International Journal of Human–Computer Interaction*, doi: 10.1080/10447318.2023.2297114.

Bozkurt, V., & Gürsoy, D. (2023). The artificial intelligence paradox: Opportunity or threat for humanity?. *International Journal of Human–Computer Interaction*, doi: 10.1080/10447318.2023.2297114.

Brinkmann, L., Baumann, F., Bonnefon, J. F. et al. (2023). Machine culture. *Nature Human Behavior,* 7(11), 1855-1868.

Calacci, D (2023). Building dreams beyond labor: Worker autonomy in the age of AI. *Intereactions Mag*, 48-51.

Capraro, V., Lentsch, A., Acemoğlu, D., et al. (2023). *The impact of generative artificial intelligence on socioeconomic inequalities and policy making*. arXiv preprint. arXiv:2401.05377.

Citron, D. K., & Pasquale, F. A. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89.

Conitzer, V., Brill, M., & Freeman, R. (2015). Crowdsourcing societal tradeoffs. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, (pp. 1213–1217). International Foundation for Autonomous Agents and Multiagent Systems.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, 538, 311-313.

de Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philos Technol,* 31, 525-541.

da Silva, J. A. T. (2021). The Matthew effect impacts science and academic publishing by preferentially amplifying citations, metrics and status. *Scientometrics*, 126, 5373-5377.

Daugherty, P. R., & Wilson, H. J. (2018). *Human + machine: Reimagining work in the age of AI*. Harvard Business Review Press.

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023). The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (EAAMO '23). Association for Computing Machinery, New York, NY, USA, Article 37, 1–23.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structure. *Digital Journalism*, 3(3), 1-18.

Didier, E. (2015). Gabriel Tarde and statistical movement. In V. Vargas (Ed), *The social after Gabriel Tarde* (pp. 299-325). Routledge.

Erdi, P. (2020). *Ranking: The unwritten rules of the social game we all play.* Oxford University Press.

Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*, 18(2), 149–156.

Frank, M. R., Autor, D., Bessen, J. E., et al. (2019). Toward understanding the impact of artificial intelligence on labor. *PNAS*, 116(14), 6531-6539.

Friedman, B. (1996). Value-sensitive design. *Interactions*, 3(6), 16–23.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, 14(3), 330-347.

Fu, R., Huang, Y., & Singh, P. V. (2020). *AI and algorithmic bias: Source, detection, mitigation and implications*, SSRN, doi:/10.2139/ssrn.3681517

Gerlich, M. (2023). Perceptions and acceptance of artificial intelligence: A multi-dimensional study. *Social Sciences,* 12(9), 502.

Harari, Y. N. (2017). Reboot for the AI revolution. *Nature*, 550(19), 324-327.

Hernandez-Ramirez, R. (2017).Technology and self-modification: Understanding technologies of the self after Foucault. *Journal of Science and Technology of the Arts*, 9(3), 45-57.

Hossain, S. Q., & Ahmed, S. I. (2021). *Towards a new participatory approach for designing artificial intelligence and data-driven technologies.* ACM, New York, NY, USA.

Hussain, S., Sanders, E. B. N., & Steinert. M. (2012). Participatory design with marginalized people in developing countries: Challenges and opportunities experienced in a field study in Cambodia. *International Journal of Design* 6, 2, 91–109.

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., & Shestakofsky, B. (2021). Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius*, 7.

Kostick-Quenet, K. M., & Gerke, S. (2022). AI in the hands of imperfect users. *NPJ Digital Medicine*, 5, 197.

Lee, M. K., Kusbit, D., Kahng, A. et al. (2019). WeBuildAI: Participatory framework for algorithmic governance. *Proc. ACM Hum.-Comput. Interact*., 3(CSCW), 181.

Lewis, K., Kaufman, J., Gonzales, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks,* 30, 330-342.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167.

Liu, Z. (2020). Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass*, 15(3), e12851.

Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, 13(5), 14-19.

Manyika, J., & Sneader, K. (2018). *AI, automation, and the future of work: Ten things to solve for*. McKinsey Global Institute.

Mayer-Schönberger, V., & Cukier, K. (2014). *Big data.* Houghton Mifflin Harcourt.

Merton, R. K. (1968). The Matthew effect in science. *Science*, 159, 53–63.

Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: Challenges and mitigation strategies. *NPJ Digital Medicine*, 6, 113.

Montesano, A. (2011). Ricardo on machinery. What matters: Technical progress or substitution of machines for circulating capital?. *History of Economic Ideas*, 19(1), 103-124.

Mu, W. (2023). How artificial intelligence affects workforces: The impact of biased recruitment and job displacement risk. *Highlights in Business, Economics and Management,* 23, 19-25.

Muller, M. J. (2009). Participatory design: The third space in HCI. In *Human-computer interaction* (pp. 181-202). CRC Press.

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C, Moukheiber, M., Khanna, A. K., et al. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, 2(6):e0000278.

Ntoutsi, E., Fafalios, P., Gadiraju, U., et al. (2020). Bias in data-driven artificial intelligence systemsan introductory survey. *WIREs Data Mining Knowl Discov*, 10(3):e1356.

Nevmyvaka, Y., Feng, Y., & Kearns, M. (2006). Reinforcement learning for optimized trade execution. In Proc 23rd *International Conference on Machine Learning* (pp. 673-680). ACM.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366, 447-453.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Books.

Özer, M. (2023a). The Matthew effect in Turkish education system. *Bartın University Journal of Faculty of Education,* 12(4):704-712.

Özer, M. (2023b). Matta Etkisi. *Uluslararası Yönetim İktisat ve İşletme Dergisi,* 19(4):974-984.

Özer, M. (2024). Başarı oyununda Matta Etkisi ve ödülün asimetrik dağılımı. *Reflektif Journal of Social Sciences,* 5(1):187-197.

Özer, M., & Perc, M. (2020). Dream and realities of school tracking and vocational education. *Palgrave Communications*, 6(1), 1-7.

Özer, M., & Perc, M. (2021). Impact of social networks on the labor market inequalities and schoolto-work transitions. *Journal of Higher Education*, 11(1):38-50.

Özer, M., Perc, M., & Suna, H. E. (2024). AI bias and the amplification of inequalities in the labor market. *Journal of Economy, Culture and Society,* 69, 159-168.

Özer, M. (2024). Potential benefits and risks of artificial intelligence in education. *Bartın University Journal of Faculty of Education,* 13(2), 232-244.

Pajarinen M., Rouvinen P., & Ekeland, A. (2015). *Computerization threatens one-third of Finnish and Norwegian employment.* ETLA Brief, 34.

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health,* 9(2), 010318.

Parkes, D. C., & Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science*, 349, 267-272.

Pasquale, F. (2011). Restoring transparency to automated authority. *Journal of Telecommunications & High Technology Law*, 9, 235-256.

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

Pelacho, M., Ruiz, G., Sanz, F., Tarancon, A., & Clemento-Gallardo, J. (2020). Analysis of the evolution and collaboration networks of citizen science scientific publications. *Scientometrics*, doi: 10.1007/s11192-020-03724-x.

Perc, M. (2014). The Matthew effect in empirical data. *Journal of the Royal Society Interface,* 11, 98.

Perc, M., Özer, M., & Hojnik, J. (2019). Social and juristic challenges of artificial intelligence. *Palgrave Communications*, 5, 61.

Piano, S. L. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities & Social Sciences Communications,* 7, 9.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology,* 20(1), 5-14.

Rahwan, I., Cebrian, M., Obradovich, N. et al. (2019). Machine behavior. *Nature*, 568, 477-486.

Reinsel, D., Gantz, J., & Rydning, J. (2018). *The digitization of the world: From edge to core*. IDC White Paper.

Rigney, D (2010). *The Matthew effect: How advantage begets further advantage*. Columbia University Press.

Schwartz, R. D. (1989). Artificial intelligence as a sociological phenomenon. *The Canadian Journal of Sociology*, 14(2), 179-202.

Shinde, P. P., & Shah, S. (2018). *A review of machine learning and deep learning applications*. 2018 Fourth International Conference on Computing Communication Control and Automation, 1-6.

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial bias in pulse oximetry measurement.

Sonnenburg, S. et al. (2007). The need for open source software in machine learning. *J Mach Learn Res*, 8, 2443-2466.

Soori, M., Arezoo, B., & Dastres, R. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics,* 3, 54-70.

Stahl, B. C. (2023). *Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems.* Scientific Reports, 13, 7586.

Strevens, M. (2006). The role of the Matthew effect in science. *Studies in History and Philosophy of Science Part A,* 37(2), 159-170.

Suleyman, M. (2023). *The coming wave: Technology, power, and the twenty-first century's greatest dilemma.* Crown: New York.

Thimbleby, H. (2003). Explaining code for publication. *Softw: Pract Experience*, 33, 975-1001.

Thompson, E. P. (1967). Time, work-discipline and industrial capitalism. *Past & Present,* 38, 56-97.

Tramer, F. et al. (2017). *Ensemble adversarial training: attacks and defences*. arxiv.org/abs/1705.07204.

Turkle, S. (1984).*The second self: Computers and the human spirit.* Simon and Schuste.

Ulnicane, I., & Aden, A. (2023). Power and politics in framing bias in Artificial Intelligence policy. *Review of Policy Research,* 40, 665-687.

Ye, F. (2017). *Scientific metrics: Towards analytical and quantitative sciences*. Science Press Beijing & Springer.

Wan, Y., & Chang, K. W. (2024). *The male CEO and the female assistant: Probing gender biases in text-to-image models through paired stereotype test*. arXiv: 2402.11089.

Wolfe, A. (1991). Mind, self, society, and computer: Artificial intelligence and the sociology of mind. *American Journal of Sociology,* 96(5), 1073-1096.

Woolgar, S. (1985). Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology*, 19(4), 557-572.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology & Human Values*, 41(1), 118-132.

Zimmer, M. (2008). *More on the 'Anonymity' of the Facebook dataset: It's Harvard College*. MichaelZimmer.org Blog.

Zuckerman, H. (1989). Accumulation of advantage and disadvantage: The theory and its intellectual biography. In Ed. C Mongardini and S Tabboni (Eds). *Robert K. Merton and contemporary sociology* (pp.153-176). New Brunswick, NJ: Transaction.