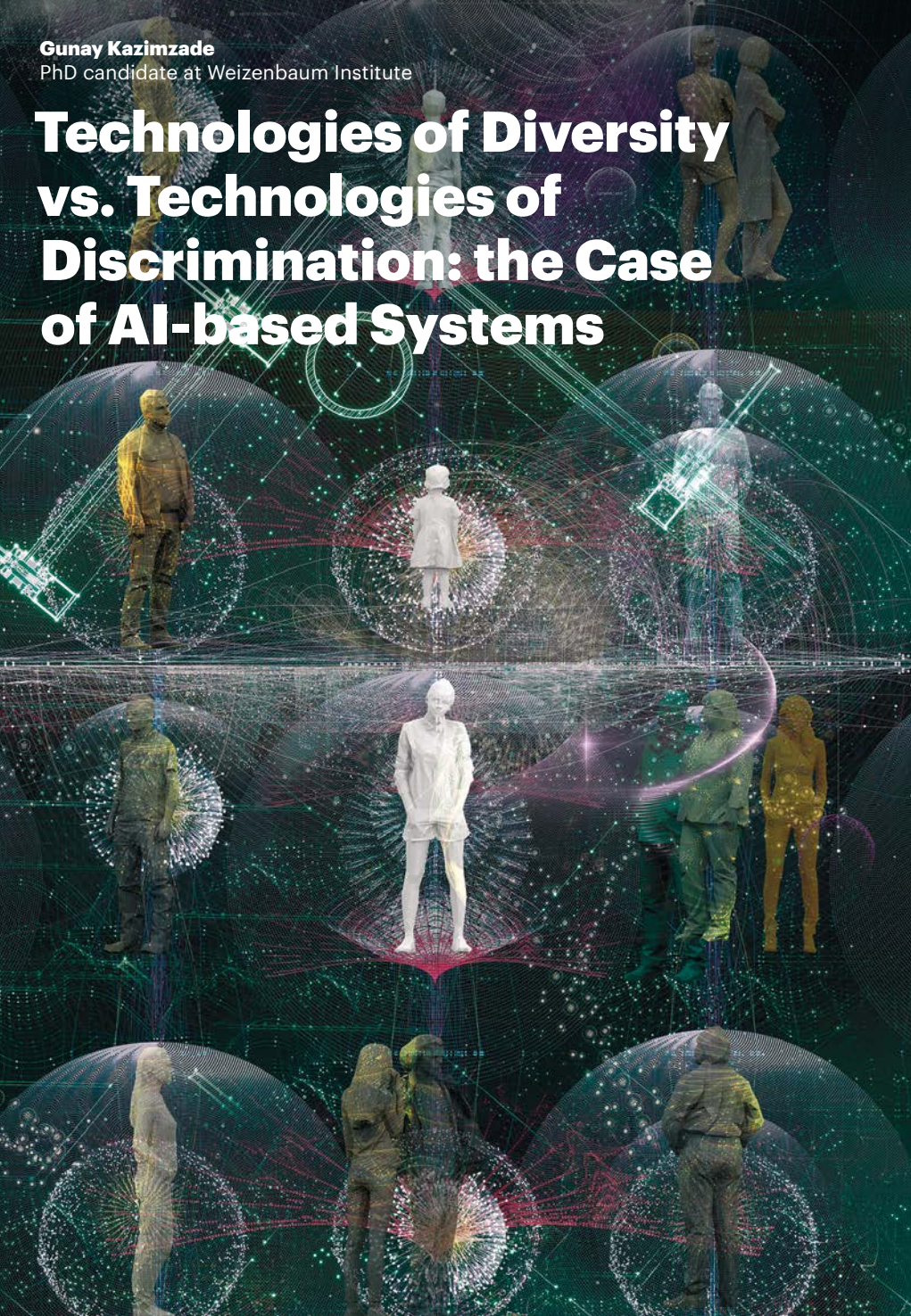
Gunay Kazimzade
PhD candidate at Weizenbaum Institute

Technologies of Diversity vs. Technologies of Discrimination: the Case of AI-based Systems



Introduction

Today, AI systems are used in a variety of domains and, from a high-level view, these technologies function as systems of discrimination: they differentiate, rank, and categorize and therefore, in some specific cases, discriminate and create inequalities in society. As the current facial recognition systems are miscategorizing people of color, women are consistently underpaid, and automatic recruitment systems are excluding female candidates for technical and leadership positions, society faces the challenge of being "categorized," "discriminated" and "unfairly judged" by intelligent systems.[1,2,3,4]

As stated by the AI Now Institute report, there is a crisis of diversity[5] in the AI sector across gender and race.[6] Authors of leading AI conferences, decision-makers, workers, and research staff of "tech giants" such as Google, Facebook and Microsoft are predominantly white and male. Also, there is no public data on trans workers or other gender minorities, as stated in the same report. Even though there is a growing concern and social focus on "fixing" diversity problems of the AI industry by approaching data quality, fair models, and inclusive design, many argue that there should be a deeper analysis of workplace cultures, power asymmetries, harassment, exclusionary hiring practices and unfair compensation that are causing people to leave or avoid working in the AI sector altogether.[7]

Therefore, it seems that the inequality problem of AI is not just a technical problem, but an issue that needs to be addressed from the interdisciplinary perspective involving different stakeholders, decision-makers, and, most importantly, civil society.

AI-based technologies are increasingly positioned in the center of our lives, developing new horizons for society. This buzzword "AI" is used to generalize technologies and systems which "imitate" human intelligence using a variety of techniques such as automatic speech recognition, image recognition, natural language processing, speech

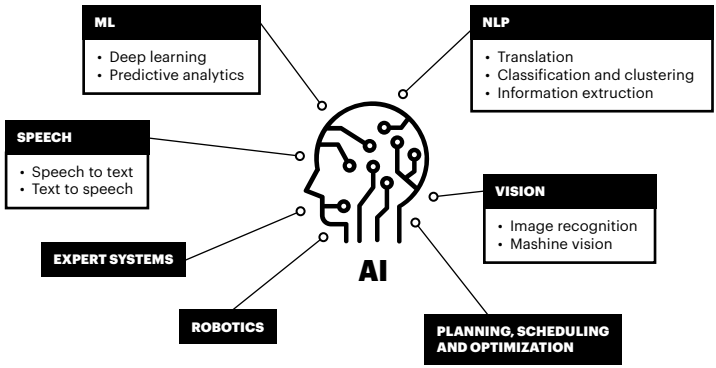


Fig. 9. Which AI-based technologies are experienceing a diversity crisis?

generation and so on. Machine learning is a subset of AI and focuses on the ability of machines to receive a huge amount of data and learn from it without being explicitly programmed. However, expert systems, which are also included under this "AI" umbrella, can operate with general programming techniques with or without machine learning algorithms. Therefore, it is important to differentiate between AI, machine learning, and other terms with regards to the scope and impact they have on the diversity and discrimination problem. In our article we focus on all these techniques and the discrimination problems caused by their implementation in a variety of domains.

Debates are ongoing on whether AI-based systems improve the quality of human life or, in contrast, increase inequalities and exclusion in society. Large-scale machine learning and deep learning techniques which enable computers to process and analyze vast amounts of data are widely used in domains such as insurance (specifically in credit scoring), loan applications, healthcare including healthcare analytics, healthcare robotics and illness diagnostics, in public safety and security (specifically in predictive policing and crime applications), in human workforce replacement and human resource management, in social media applications, games, digital entertainment services and in theeducational domain for teacher robots,

children-robot interaction, intelligent tutoring systems, online learning and learning analytics.[9,10,11,12]

In specific use-cases, however, these socio-technical systems bring unfair, unethical, and discriminating results. Report by the AI Now Institute states that "Systems that use physical appearance as a proxy for character or interior states are deeply suspect, including AI tools that claim to detect sexuality based on a picture of someone's head, predict 'criminality' based on facial features, or assess worker competence via 'micro-expressions.' Such systems are replicating patterns of racial and gender bias in ways that can deepen and justify historical inequality. The commercial deployment of these tools is a cause for deep concern."[13]

The scandal involving Amazon's "sexist" AI-based recruitment tool which "learned" to eliminate female candidates was brought to public attention and the company itself in 2018. The reason behind the unfair judgements made by the system was its use of historical data that captured decisions made by human recruiters in the past 10 years. During that period very few women were hired to leadership and technical positions; therefore, the system trained on that data learned to imitate the biased decisions made by human workers at the company. After the scandal went viral, the company decided to edit the program in order to neutralize gender features; however, there is still no guarantee that the recruitment systems would not correlate other features with the candidates' gender attributes.

Timnit Gebru, who studies algorithmic bias at Microsoft, emphasizes[14] concerns about how deep learning could reshape the insurance market; minority and underrepresented groups may be discriminated against due to a higher volume of traffic collisions in more densely populated zones where they are more likely to live. A deep-learning program could "learn" that there is a correlation between belonging to a minority group and a higher volume of traffic collisions and use this to build a model with prejudices against people of color, for instance. In

this case, that insurance system would have developed a racial bias.

Machine-learning algorithms in AI-based systems are currently applied in the healthcare industry to analyze high volumes of data to improve decision-making, guide treatment decisions and improve efficiency. Such data, collected over several years, can reflect historical biases against vulnerable populations. It leads to potential promotion of further bias, leading to disparity in the healthcare industry.

Predictive policing algorithms are becoming immensely popular in cities across the US, as well as in other countries. Many researchers and privacy scholars are concerned about critical consequences of decisions made by such systems, since they have the potential to reinforce racial and cultural biases. "Police in America is systematically biased against communities of color," according to New York Civil Liberties Union legal director Christopher Dunn told Fast Company. "Any predictive policing platform runs the risks of perpetuating disparities because of the over-policing of communities of color that will inform their inputs. To ensure fairness, NYPD should be transparent about the technologies it deploys and allow independent researchers to audit these systems before they are tested on New Yorkers."[15]

With the boom of smart technologies, social media platforms have become trusted spaces to share personal information, photos, activities, and discussions of topics such as politics, religious views, and other sensitive subjects. In order to operate at a larger scale, these platforms are applying AI-based techniques in filtering and targeting methods in recommendation systems for movies, music, and news channels, as well as news feed generation on social media platforms. They are manipulated with respect to the users' demographic information, gender, age and browsing history, thus providing information which fits within their existing "bubble" world view and explicit interaction with those who share that view. Over

time, the biases and prejudices of those filter bubbles are reinforced and distributed within these communities. The issue is the same for non-traditional interfaces including Amazon Alexa and Microsoft Cortana. Growing numbers of users are experiencing these interactions manipulated by smart algorithms and there is a danger of these technologies limiting our choices and interactions without us even realizing it.

Considering the fast-paced evaluation of algorithmic decision processes it is likely that they will be increasingly affecting society in the coming years. It is vital that civil society speaks up about issues such as bias and discrimination in AI-based systems, as well as strategy, vision, and action plans to overcome these issues.

Possible development directions and desirable future

First we must consider how this can be done. What developments could society face with the exponential growth of data and implementation of AI-based systems in different domains, in particular socio-technical systems?

The first step towards solving the discrimination problem in AI requires the application of gender- and cultural-sensitive guidelines for fair data collection, data handling, design, and implementation layers of the AI-pipeline. Moreover, it is vital for each level of society, including governmental organizations, businesses, NGOs and educational institutions, to follow these guidelines and apply them in their own specific domains.

The other direction of development is in solving the inequality problem by applying emerging technologies for the needs of civil society. Data-driven applications trained on incomplete datasets, which only capture limited cultural or geographic groups, may produce results biased against other groups that were not captured in these datasets. This happens due to a lack of availability

of quality data in these geographic locations. For instance, in the dataset presented by UK Biobank which aimed to genotype 500,000 individuals, ethnic minorities were significantly underrepresented, including Black (by a third), Chinese (by more than a third) and Indian and Pakistani (by more than half). White British participants make up 94.6% of Biobank samples, compared to 91.3% of the general population. This sample has a dramatic impact on medical diagnostics, creates a bias and increases the risk of wrong diagnoses in the underrepresented groups.[16] AI developers are targeting European and US populations due to the lack of quality data representing other populations.[17] Therefore, open data initiatives in marginalized communities may provide a unique opportunity to include underrepresented groups in the agenda of technological solutions aimed at solving cultural diversity problems.

The use of machine learning and AI-based technologies in the educational domain is one of the least discussed applications with respect to the role of emerging technologies in reducing social inequality. However, it shows promise in overcoming problems of social inequality caused by emerging technologies. With the current rapid development of technology, decision making, technology development and data collection are manipulated by a small elite. Thus, there is an opportunity to distribute this knowledge and power among all layers of the society. This is possible by educating the next generation of female tech leaders, teaching state-of-the-art technologies such as artificial intelligence and machine learning at an early age, teaching interdisciplinarity, promoting intercultural cooperation and diversity as well as conscious and unconscious human biases reflected on technologies impacting the society.

Civil society should play an essential role in this case by understanding and adapting data-driven technologies and privacy, and their political and economic influences, as well as new opportunities and risks that these technologies bring.

With respect to such issues as social media profiling, political manipulation and discriminating decision-making systems, civil society could serve as a bridge between society and policy-makers and technology reinforcers in bringing the communities they serve to the table and including underrepresented groups in development processes.

The most relevant role of civil society here is in understanding the dangers of the biases caused by AI systems and how they may affect the issues they implement and the people and communities they serve. The goal of civil society organizations could be set towards raising awareness of companies and organizations implementing new algorithms on challenges such as fairness, transparency, and accountability; the same applies to policymakers, responsible for forming new laws and regulations, designed to govern these technologies, as well as monitoring and analyzing the impact and consequences of the implemented strategies and standards.

What could go wrong?

Without proper safeguarding, AI-based systems may bring negative consequences to society by creating an authoritarian and centralized way of manipulating, filtering, and discriminating underrepresented groups in society. Inequality of access and geographic underrepresentation could be applied by manipulating training data and machine learning models in critical cases such as employment. This could lead to the use of these technologies for distributing uneven power among society, power of distribution and creating "disconnected" bubbles in society.[18] AI technologies may also bring privacy-related issues with respect to personal data, fake-news and political manipulation through targeting and filtering on social media; these are the ultimate risks of the "divide and conquer" approach that threatens democracy.[19]

The centralization or decentralization of power within these technologies can have consequences concerning societal equity and equality if and when such technologies are used as a tool to manipulate, govern, and direct further development of society. In this sense, for instance, China's social credit system has been compared to the Black Mirror TV series, Big Brother reality show, and other dystopian future science fiction narratives. "What's really scary is there's nothing you can do about it. You can report to no one. You are stuck in the middle of nowhere," says one of the black-listed journalists from China who was "tagged" as "not qualified" to buy a plane ticket and banned from travelling by certain train lines, buying property, or taking out a loan.[20]

Unknowns

Extensive discussions on the topics of transparency, fairness and accountability of the algorithms and technologies used to impact society are ongoing, although they are not fully incorporated in the design and implementation of these technologies. Not all machine learning and deep learning algorithms can explain their decisions, and most of the data used to develop data-driven systems is biased and does not capture the entire population it is aimed at. How will "black box" algorithms be governed? How will AI biases impact the people AI is aimed at? Who is responsible for governing all these technologies? We are yet to find answers to these questions. There are, however, ongoing initiatives of organizations such as the AI Now Institute, Alan Turing Institute and Leverhulme Institute for the Future of Intelligence which are raising awareness on approaching crucial problems of discrimination and exclusion problems in data-driven systems, as well as the implementation of new technological solutions for eliminating the negative impact of issues concerning AI strategies and policies introduced by different decision-makers and commissions.

Weak signals regarding the role of emerging technologies

Although we discuss bias and discrimination caused by AI-enabled technologies, it is possible to use such technologies to detect biases occurring at the different stages of the technology development lifecycle. It is one of the directions which has a slow but important impact on overcoming inequality problems in society.

For instance, a team of US researchers has developed an AI tool for detecting bias based on race and gender of job or university applicants. The system is trained on a vast volume of data and makes recommendations on hiring female candidates, if they have been underrepresented in specific positions or faculties for a long time.

These kinds of technologies can be used and governed by civil society organizations, as it is the direct responsibility and aim of these organizations to measure and mitigate biases and discrimination in socio-technical systems.

doi: 10.24412/cl-35945-2021-1-146-157

Endnotes

1 O'Neil C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (1st ed.). London: Penguin.
2 Rice L., Swesnik D. (2013). Discriminatory Effects of Credit Scoring on Communities of Color. *45 Suffolk University Law Review*. no. 935. p. 32.
3 Whittaker M., Crawford K., Dobbe R., Genevieve F., Kaziunas E., Varoon M., West S.M., Richardson R., Schultz J., Schwartz O. (2018). *AI Now Report 2018*. AI Now Institute, New York University, 2018.
4 Buolamwini J., Gebru T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability, and Transparency*.
5 *Diversity* – a broad sociological and moral concept, that assumes respect and acceptance of individual and group differences such as race, ethnicity, sexual orientation, age, political and religious views, etc. – ed. note.
6 West S.M., Whittaker M., Crawford K. (2019). *Discriminating Systems: Gender, Race and Power in AI*. AI Now Institute. URL: https://ainowinstitute.org/discriminatingsystems.pdf (retrieval date 25.08.2020).
7 West S.M., Whittaker M., Crawford K. (2019). *Discriminating Systems: Gender, Race and Power in AI*.
8 Source: Deloitte Insights. URL: deloitte.com/insights.
9 O'Neil C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.
10 Rice L., Swesnik D. (2013). *Discriminatory Effects of Credit Scoring on Communities of Color*.
11 Buolamwini J., Gebru T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*.
12 Whittaker M., Crawford K., Dobbe R., Genevieve F., Kaziunas E., Varoon M., West S.M., Richardson R., Schultz J., Schwartz O. (2018). *AI Now Report 2018*.
13 Olteanu A., Castillo C., Diaz F., Kiciman E. (2016). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *SSRN Electronic Journal. Frontiers in Big Data*. no. 2:13. 20 December 2016. URL: http://dx.doi.org/10.2139/ssrn.2886526 (retrieval date 25.08.2020).
14 Gebru T., Morgenstern J., Vecchione B., Vaughan J.W., Wallach H., Hal Daumeé III, Crawford K. (2018). *Datasheets for Datasets*. arXiv.org, 9 July 2018.. URL: http://arxiv.org/abs/1803.09010 (retrieval date 25.08.2020).
15 O'Neil C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.
16 Swanson J.M. (2012). The UK Biobank and selection bias. *The Lancet*. 14 July 2012. URL: https://doi.org/10.1016/S0140-6736(12)61179-9 (retrieval date 25.08.2020).

17 Olteanu A., Castillo C., Diaz F., Kiciman E. (2016). *Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries*.

18 O'Neil C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.

19 Bourdieu P. (1989). Social Space and Symbolic Power. *American Sociological Association: Sociological Theory*. no. 7(1). pp.14-25. URL: https://doi.org/10.2307/202060 (retrieval date 25.08.2020).

20 Matsakis L. (2019). How the West Got China's Social Credit System Wrong. *Wired*. 29 July 2019. URL: https://www.wired.com/story/china-social-credit-score-system/ (retrieval date 25.08.2020).