NO BLACK BOX: PROMOTING INCLUSION AND DEMOCRACY IN THE AGE OF ARTIFICIAL INTELLIGENCE

NO BLACK BOX: PROMUOVERE L'INCLUSIONE E LA DEMOCRAZIA NELL'ERA DELL'INTELLIGENZA ARTIFICIALE

Monica Di Domenico[1]
Università degli Studi di Salerno
modidomenico@unisa.it

Giuseppina Rita Mangione
Indire
g.mangione@indire.it

Elsa Maria Bruni
Università degli Studi "G. d'Annunzio" Chieti-Pescara
elsa.bruni@unich.it

Double Blind Peer Review

Citazione
Di Domenico, M., Mangione, G.R., & Bruni, E.M. (2024). No black box: promoting inclusion and democracy in the age of artificial intelligence. *Giornale Italiano di Educazione alla Salute, Sport e Didattica Inclusiva*, 8(2), Edizioni Universitarie Romane.

Doi:
https://doi.org/10.32043/gsd.v8i2.1144

gsdjournal.it
ISSN: 2532-3296
ISBN 978-88-7730-493-3

ABSTRACT
This work dissects how the ascendant role of AI in adjudicating citizenship status is leading a new kind of citizenship shaped by digital interactions: the "Algorithmic Citizenship". The analysis delves into the threat of algorithmic discrimination and of algorithmic historical revisionism. The work emphasizes the importance of explainability and transparency in algorithms and it examines the challenges posed by their "black box" nature.

Questo lavoro analizza come il ruolo crescente dell'IA nell'attribuzione dello status di cittadino stia generando una nuova forma di cittadinanza plasmata dalle interazioni digitali: la "Cittadinanza Algoritmica". L'analisi approfondisce la minaccia della discriminazione algoritmica e della revisione storica algoritmica. Il lavoro sottolinea l'importanza della interpretabilità e trasparenza degli algoritmi ed esamina le sfide poste dalla loro natura di "black box".

KEYWORDS
Algorithmic Citizenship, Artificial Intelligence, Bias, Education, Critical Thinking
Cittadinanza algoritmica, Intelligenza Artificiale, Pregiudizio, Educazione, Pensiero Critico

Received 30/04/2024
Accepted 15/06/2024
Published 24/06/2024

[1] The article is the result of the joint work of the authors, who together planned and created the structure and contents. Given the above, G.R. Mangione wrote the introduction, E.M. Bruni the conclusions and M. Di Domenico the paragraphs 2, Mitigating Bias in Machine Learning: Preserving the Integrity of the Past and 3, No Black box: Achieving Transparency and Explainability in AI.

Introduction

Historically, the acquisition of citizenship has been tied to two fundamental principles: *Jus Soli*, which grants citizenship to individuals born within the borders of a state regardless of their parents' nationality, and *Jus Sanguinis*, which attributes citizenship based on the citizenship of their parents. However, the complexities of the modern world, characterized by migration, globalization, and rapidly evolving technologies, challenge this traditional view of a stable and absolute notion of citizenship.

In the online world, our rights and identities, no longer solely linked to a physical location, intertwine with our digital identities, composed of a multitude of online information. These digital identities play an increasingly important role in defining our rights and our relationship with states, financial institutions, and society. In this context, the neologism “Jus Algoritmi”, coined by Cheney-Lippold, describes an emerging form of citizenship generated by the surveillance state that operates through identification and categorization. This results from the extensive use of software in decision-making processes regarding an individual's citizenship status (Cheney-Lippold, 2016).

Latour recognizes society as being formed by networks that involve both human and non-human actors on an equal footing. This suggests an analysis of the influence of non-human actors, such as algorithms and digital platforms, on the emergence and manifestation of group identities. Digital platforms act as crucial nodes in emerging social networks, facilitating communication, the sharing of experiences and the construction of virtual communities that transcend geographical and cultural boundaries. In this context, communication technologies are no longer a neutral medium for transmitting information and their use contributes to the formation of new collective identities and mass identification (Latour, 2007).

The concept of “Algorithmic Citizenship” expands the perspective of “Jus Algoritmi”, the latter referring to the right to citizenship determined by an algorithm that assumes a role traditionally reserved for legal procedures and institutions. *Algorithmic Citizenship* refers to an individual's participation in society through digital interactions, where digital platforms, guided by algorithms, shape social experiences and civic participation. This concept incorporates the notion that citizenship is no longer defined solely by geographical boundaries but also by digital participation and online visibility.

The very idea of Algorithmic Citizenship raises the possibility of *Algorithmic Discrimination*, a phenomenon in which algorithms can amplify and perpetuate

existing injustices, relying on training data that reflects social biases (Bridle, 2016). To analyze the concept of Algorithmic Discrimination, it is necessary to refer to the concepts of *Bias* and *Explainability* of algorithms.

Bias refers to the presence of prejudices in the data on which algorithms are trained. When the data used for training implicitly or explicitly contains cultural, social, or other biases, algorithms can inherit, perpetuate, and multiply these biases, influencing their decisions and predictions.

1. Algorithmic Discrimination and Bias: A Critical Examination

Numerous examples and scientific studies have highlighted the complex issues related to algorithmic bias and discrimination. All Generative Artificial Intelligence systems are based on powerful forms of machine learning, where algorithms *learn* to predict particular outcomes from patterns and structures in vast datasets. Neural networks are a type of machine learning algorithm that can be used to learn the relationships between inputs and outputs. When a neural network is trained on a set of data, it is able to generate new data that is similar to the data it was trained on. If the training data is biased or incomplete, so too will be the results the system arrives at, discriminating against certain individuals or groups, amplifying and perpetuating injustices and prejudices of the past.

The American criminal justice system serves as a stark example in this regard: when probation committees in the United States began using data to predict the risk of recidivism, they encountered a century of racism embedded in the data. The stories of bias in the American criminal justice system are reflected in the data used to train machine learning algorithms, and these algorithms can therefore reproduce and amplify those patterns of injustice. Here are the most remarkable cases:

- An experiment conducted by ProPublica in 2016 found that the risk prediction software used in many American courts is heavily biased against African Americans. The experiment discovered that the software was more likely to predict that African Americans would re-offend, even when there was no evidence to support this claim. This led to a disproportionate number of African Americans being incarcerated (Angwin et al., 2016).
- In 2017, a research conducted by MIT found that facial recognition systems are more likely to misidentify people of color as suspects. Consistent with previous research, this study also shows that facial recognition systems were 35% more likely to misidentify African Americans as suspects than

whites. This led to a disproportionate number of people of color being falsely arrested. (Buolamwini, J., & Gebru, T., 2018).

- An experiment conducted by the New York City Department of Correction in 2018 found that the prison cell assignment system was heavily biased against African Americans. The experiment found that African Americans were more likely to be assigned to overcrowded and unsanitary cells compared to whites. This led to a disproportionate number of African Americans contracting illnesses while in prison (Peck, J., 2018).

Benjamin, in "Assessing Risk, Automating Racism," provides a critical perspective on the risk of automating racism through the use of algorithms in decision-making contexts, including those related to public safety (Benjamin, R., 2019). A comprehensive analysis of how search engine algorithms can perpetuate racial and gender stereotypes, promoting a distorted representation of society, is conducted by Safiya Umoja Noble, author of "Algorithms of Oppression" (Noble, S. U., 2018).

It is necessary to delve into the various types of bias to deepen the link between bias and algorithmic discrimination. In relation to the constant demand for equity in algorithmic outcomes, closely linked to the concept of cultural, social, economic, physical, cognitive, and gender diversity, Rivoltella and Panciroli (2023) analyze examples of bias in outcomes and how they can influence user decisions and the feedback cycle. Referring to Suresh and Guttag's proposal (2021), they identify four main types of bias:

- "Measurement Bias": concerns the methods of selection, use, and measurement of particular characteristics;
- "Omitted Variable Bias": involves the exclusion from the model of one or more important variables;
- "Representation Bias": stems from the sampling method of a population during data collection;
- "Aggregation Bias": occurs when erroneous conclusions are drawn about observed individuals within an entire population.

Regardless of the types of bias, it is important to consider that sexist and racist assumptions, even based solely on underrepresentation, are ingrained in industrial culture and perhaps even more so in the subculture of the technology industry. Therefore, biases may be mitigated through specific techniques but not eliminated entirely.

2. Mitigating Bias in Machine Learning: Preserving the Integrity of the Past

The complex interplay between the concepts of bias, algorithmic discrimination and historical revisionism deeply influences our understanding of the past and the formation of collective memory. It is crucial to balance bias correction with respect for historiographical research and the preservation of historical integrity, avoiding manipulations that could inappropriately distort historical narratives (Crawford, K., & Calo, R., 2016; Benjamin, R., 2019).

Involving diverse perspectives in design, including ethics experts, educators, sociologists and historians, can contribute to identifying and mitigating biases more comprehensively (Obermeyer et al., 2019). Careful analysis of training data is fundamental to identifying and understanding present biases.

Bias mitigation techniques introduce diversity into training data, reflecting cultural, ethnic and gender diversity (Buolamwini, J., & Gebru, T., 2018). Transparent regulations for algorithm design and implementation are suggested to ensure understandable and assessable decisions (Diakopoulos, N., 2016).

Another strategy is bias correction during the learning process using techniques that seek to balance outcomes to avoid unfair discrimination (Chouldechova, A., 2017). Techniques such as oversampling, undersampling, or generating synthetic data, artificially created or modified to extend or improve existing data (Koh, P.W., & Liang, P., 2017), are involved in an initial pre-processing approach, which implies identifying and correcting biases in data before model training. Responsible use of synthetic data is particularly critical as the quality and representativeness of such data directly influence model effectiveness.

A second approach involves careful model selection methods favoring fairness, such as those based on group or individual fairness. For example, Kamiran and Calders (Kamiran, F., & Calders, T., 2012) proposed a method to select classifiers that achieve demographic parity, fairly distributing positive and negative outcomes among different demographic groups.

A post-processing approach involves regulating the output of AI models to remove bias and ensure fairness. For example, post-processing methods have been proposed to adjust model decisions by equitably distributing false positives and false negatives among different demographic groups (Zhang, B.H., Lemoine, B., & Mitchell, M., 2018).

These approaches promise to mitigate bias in AI but come with limitations. For example, adjusting model predictions to ensure fairness may involve trade-offs between different types of bias, with potential unintended consequences on outcome distribution among different groups (Kleinberg, J., et al., 2018).

Ultimately, the scientific community underscores the importance of a balanced approach, a holistic strategy, from ensuring diversity in data to selecting transparent models and critically post-processing generated results.

3. No Black box: Achieving Transparency and Explainability in AI

A generative artificial intelligence algorithm is considered a *black box* when its internal logic or decision-making process leading to results are not easily interpretable or understandable to humans, and even opaque to the original programmers (Bornstein, S., 2018). While humans may be required to account for and justify decisions that appear to be biased, machines may not be able to provide such explanations, nor their creators.

The reasons why machine learning models operate as black boxes are varied: Firstly, the massive training dataset can make it difficult (and costly) to explain to a human how the model was able to learn the relationships between inputs and outputs. Secondly, machine learning models are often based on neural networks, which are highly complex mathematical systems capable of learning nonlinear relationships between inputs and outputs, but it is difficult to explain how these relationships are learned.

The black box nature of AI algorithms represents a significant problem especially in contexts such as in healthcare, law, or finance, where decisions can have a significant impact on people's lives, and understanding the reasons for a particular decision is essential. The ability to clearly and understandably explain the decision-making process of AI algorithms, providing a rationale or justification for their predictions or actions, is referred to as *Explainability*. Many studies highlight the importance of addressing the issue of the *black box* in AI algorithms to ensure responsible and fair use of such technologies. In particular, Lipton's work (2018) has helped clarify the concept of Explainability by identifying more than one definition, each linked to goals and context, distinguishing between:

- *Local Explainability*: The ability to understand the decisions made by the model for a single data point;

- *Global Explainability*: The ability to understand the generalizations made by the model across a dataset;
- *Causal Explainability*: The ability to understand the causal mechanisms the model is learning.

Furthermore, the author advocates for the importance of balancing explainability and accuracy while considering the needs of different stakeholders in the design and implementation process of machine learning systems. Explainability may seem like a top priority for all language models, but in a market context, it may conflict with other aspects of AI, such as accuracy or computational efficiency. More interpretable models may sacrifice some degree of accuracy or require more computational resources compared to more complex models. Therefore, in balancing the need to explain AI decisions with the demand for efficient performance, different contexts may lean towards performance at the expense of transparency.

The fundamental importance of explainability lies in promoting transparency and accountability of AI algorithms and in creating trust and social acceptability among all users, especially in educational settings. Explaining how an algorithm reaches its predictions or decisions helps dispel doubts and provide an intelligible justification for its actions, thereby increasing confidence in the system. However, explainability and transparency are industry-specific concepts, what is transparent to an AI researcher may not be so for an end user. Therefore, system transparency should be evaluated from the perspective of the intended end users, so that they have ultimate control over when and how to use the tools, with the machine assuming a supportive role.

Key elements in this direction include model openness, open science, and open-source code. Opening up parameters of AI models is seen as a way to democratize such capabilities. The open-science approach provides a rich framework for transparently documenting the development process of such models and improving understanding of the ethical and fairness aspects of new technologies. Open-sourcing of source code is considered a means to enable effective use of AI tools, coupled with necessary knowledge transfer for full exploitation of these advanced resources. Finally, the importance of collaboration between experts and beneficiaries is emphasized, as well as the adoption of low-code interfaces to simplify interaction between users and complex AI systems.

Conclusions: The Role of Critical Thinking in Achieving Inclusive AI

We often assume that machines are inherently objective, that they cannot help but analyze data without bias or malice. This partly stems from their nature as data-driven and algorithmic systems, which may appear as cold and impersonal mathematical processes. However, as widely argued, it is crucial to recognize that fairness is an outcome of their design and implementations. It becomes evident how the possibility of free access to datasets and language models used in Generative Artificial Intelligence systems is an indispensable condition, ultimately, of democracy, and therefore fundamental in an educational context.

Regarding the risk of bias, it is important to emphasize that inclusion, as an educational principle, is still relatively young in its evolution, and the datasets used to train AI systems may reflect biases and past practices that do not align with the current conceptualization of inclusion. For example, historical data used to train machine learning algorithms may contain discriminatory information regarding students with disabilities, students from ethnic minorities or disadvantaged groups. This data can negatively influence AI decisions and recommendations in education, creating disparities and discrimination instead of promoting an inclusive environment.

Given the inherent bias and opacity of Artificial Intelligence, education in critical thinking emerges as a fundamental element in the context of AI, as the increasing complexity and pervasiveness of such systems require informed and aware algorithmic citizenship. Critical thinking, understood as the ability to objectively analyze information, identify and evaluate arguments, becomes crucial in the context of AI, where decisions can be automated through complex algorithms. AI literacy goes beyond mere technical understanding and involves the training of individuals capable of questioning and understanding the ethical, social, and cultural implications of automated systems. In this context, education in critical thinking plays a key role in instilling the ability to interrogate training data, identify implicit biases and understand the limitations of algorithms. Individuals trained in critical thinking (Beatini, V., et al., 2024) (Di Tore, S., et al., 2020) are better equipped to actively participate in the AI development process, contributing to mitigating ethical risks and ensuring that such systems are designed and implemented responsibly. Furthermore, critical awareness allows users to understand the decisions made by algorithms, recognize any distortions in the presentation of information and exercise informed control over the use of AI-based technologies. The intersection between education in critical thinking and AI thus

represents an essential connection to develop a fair, aware digital society capable of addressing emerging challenges related to decision automation.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias in criminal justice. *ProPublica*

Beatini, V., Cohen, D., Di Tore, S., Pellerin, H., Aiello, P., Sibilio, M., & Berthoz, A. (2024). Measuring perspective taking with the "Virtual Class" videogame: A child development study. *Computers in Human Behavior*, 151, 108012.

Benjamin, R., Assessing risk, automating racism. *Science*, 2019. 366(6464): p. 421-422

Bornstein, S. (2018). Antidiscriminatory algorithms. *Ala. L. Rev., 70*, 519.

Bridle, J. (2016). Algorithmic citizenship, digital statelessness. *GeoHumanities*, 2(2), 377-381.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Cheney-Lippold, J. (2016). Jus Algoritmi: How the national security agency remade citizenship. *International Journal of Communication*, 10, 22.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2), 153-163.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, 538(7625), 311-313.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62.

Di Tore, S., Aiello, P., Sibilio, M., & Berthoz, A. (2020). Simplex didactics: promoting transversal learning through the training of perspective taking. *Journal of e-Learning and Knowledge Society*, 16(3), 34-49.27.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1), 1-33.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1), 237-293.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885-1894). PMLR.

Latour, B. (2007). Reassembling the social: An introduction to actor-network-theory. Oup Oxford.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.

Panciroli, C., & Rivoltella, P. C. (2023). Can an algorithm be fair?: intercultural biases and critical thinking in generative artificial intelligence social uses. *Scholé: rivista di educazione e studi culturali: LXI, 2, 2023*, 67-84.

Peck, J. (2018). New York City jails have a racial bias problem. *The New York Times*

Suresh, H., & Guttag, J. (2021, October). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340).