Is the Common Approach used to Identify Social Biases in Artificial Intelligence also Biased?

Ana Bucchi[1], Gabriel M. Fonseca[1]

[1] Centro de Investigación en Odontología Legal y Forense (CIO), Facultad de Odontología, Universidad de La Frontera 4811230, Chile.

Abstract

Here, we ask whether the most common approaches used to identify demographic biases in artificial intelligence (AI) are also biased. We conducted a Scoping Review of papers indexed in Scopus and WoS on biases in a particular AI application (face recognition). Fourteen original articles met our inclusion criteria. Of these, the vast majority (13) used an *a priori* approach to identify bias, i.e., they started from a known background in which social groups were subject to low accuracy by the algorithms. Only one study found bias *a posteriori*, i.e., they examined the results without underlying assumptions about the discriminated groups. Remarkably, this single article identified that it was workers who suffered the negative effects of face recognition, a social segment not analyzed by any study using an *aprioristic* approach. Of the *aprioristic* studies, 79% examined skin color and ethnicity, 50% analyzed gender, and two (14%) studied age. Only two articles analyzed bias on-the-ground, while most focused on experiments. We argue that the almost exclusive use of the common approach (*aprioristic* and experimental designs) to identify systematic errors is a methodological bias. This precludes knowledge of other discriminated social groups or even biases towards humanity as a whole that have never been identified (deep-rooted biases), since their awareness depends on the historical context. To better describe AI models, we believe that eXplainable Artificial Intelligence (xAI) tools should work together with *a posteriori* bias identification strategies and the measurement of their direct effects on citizens' lives.

Keywords

Computer Vision, methodological bias, awareness, *a posteriori*.

1. Introduction

It is well known that AI systems embody human bias towards certain demographics [1], [2]. Precisely, preventing injustice and discrimination can be facilitated using xAI tools [3]. xAI produces more explainable models and enables humans to understand, trust, and effectively manage the new generation of artificial intelligence [4]. However, because the identification and mitigation of biases in AI can only be ultimately performed by humans, it is clear that the systematic errors whose recognition is facilitated by xAI are those that are conscious or easily accessible and recognizable by a human [5]. In this study, we focused on whether there are patterns in the way we approach AI biases that prevent the recognition of discriminated social groups, which should be considered by users of xAI tools. To this end, we conducted a scoping review to learn how social biases are identified and analyzed in a specific AI application (face recognition).

2. Material and Method

A scoping review was conducted following the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA ScR) [6].

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

ana.bucchi@ufrontera.cl (A. Bucchi)

The electronic search was performed using the concepts (("bias*) AND ("fac* recognition" OR "fac* verification" OR "fac* identification") AND ("artificial intelligence" OR "machine learning" OR "deep learning")) in two indexed databases (Scopus and WoS). Only articles published in English and with online accessible full texts were included. To be included, articles had to specifically address biases that occur when AI reflects discrimination towards certain social groups [1]. They also had to focus on automatic face recognition systems, which involve detecting a face in a photo or video and identifying or verifying who that person is [7]. Reviews and letters to the editor were excluded.

One reviewer (A.B.) conducted the search and each article was analyzed according to the following variables:

1. Types of knowledge: Whether biases were identified *a priori* or *a posteriori*. The former refers to studies that, as an antecedent to the research, selected a social category (e.g., gender and skin color) and evaluated the accuracy of face recognition systems according to them. In contrast, by *a posteriori* study, we refer to studies that start without assumptions about which social group is discriminated against or whether there was discrimination at all, but rather evaluated the results of the AI for patterns in the errors.

2. Research design: whether biases were identified in the field or in a controlled experiment. The former studies examined recognition systems in practice, while the latter focused on analyzing databases, algorithms, and predictions in controlled experiments.

The search was conducted on April 13, 2023, and the identified articles were screened, evaluated, and included between April 14, 2023, and May 15, 2023. The analyzed variables were recorded using an Excel spreadsheet (Microsoft Excel).

3. Results

The literature search yielded 178 articles. Following screening of titles and abstracts and after establishing eligibility (i.e., whether they were related to the study objectives), 14 articles were included in this qualitative synthesis. Figure 1 shows the flow of article selection from identification to inclusion. Table 1 shows all the articles included according to the variables assessed in this review.

3.1. Types of Knowledge

Of the 14 included articles, the vast majority (13) were *aprioristic* studies, while only one did not start with underlying bias assumptions (*a posteriori* study) [8]. Of the *aprioristic* studies, 79% studied skin color and ethnicity, 50% analyzed gender, and two (14%) studied age (Table 1). The *a posteriori* study analyzed UBER drivers' perceptions of their facial verification system and found that drivers dynamically innovate and create numerous strategies to fix the verification errors: they tilted their face, moved it closer to the light, removed their hat and glasses, changed their hairstyle, bring their faces outside, placed it in front of headlights, and took it into well-lit restrooms at gas stations or under bright lamps in dark parking lots.

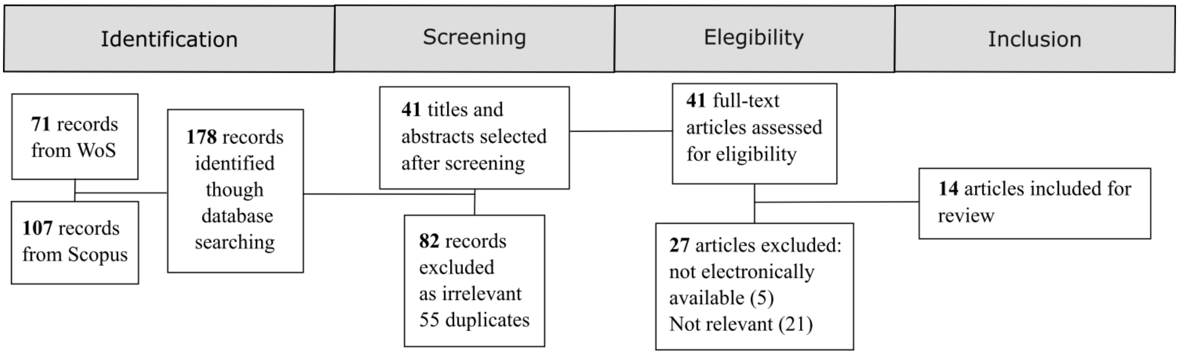
Identification
Screening
Elegibility
Inclusion
71 records from WoS
107 records from Scopus
178 records identified though database searching
41 titles and abstracts selected after screening
82 records excluded as irrelevant 55 duplicates
41 full-text articles assessed for eligibility
27 articles excluded: not electronically available (5) Not relevant (21)
14 articles included for review

Figure 1: Flow of selection process for eligible studies

3.2. Research Designs

Two articles showed the biases associated with facial recognition systems in the field [8], [9]: Watkins' study [8] analyzed the use of Uber's verification system in New York City (USA) and Toronto (Canada) through semi-structured interviews to find out workers' perceptions of UBER's verification system, while Johnson et al. [9] analyzed 1136 cases of arrests in the USA using facial recognition and their relationship with black or white inmates. The other articles used experimental designs to test whether the databases, algorithms or predictions produced different accuracy according to certain social categories (all determined *a priori*).

Table 1
Articles included in the review and variables analyzed

Reference	Type of knowledge	Research Design
Watkins [8]	*A posteriori* (workers)	On the ground
Albiero et al. [10]	*A priori* (gender)	Experimental
Franco et al. [11]	*A priori* (ethnicity and gender)	Experimental
Georgopoulos et at. [2]	*A priori* (kinship, gender and age)	Experimental
Georgopoulos et al. [12]	*A priori* (skin color, gender and age)	Experimental
Celis and Rao [13]	*A priori* (skin color)	Experimental
Johnson et al. [9]	*A priori* (race)	On the ground
Coe and Atay [14]	*A priori* (race)	Experimental
Pagano et al. [15]	*A priori* (gender)	Experimental
Wang et al. [16]	*A priori* (skin color)	Experimental
Serna et al. [17]	*A priori* (ethnicity and sex)	Experimental
Jiang et al. [18]	*A priori* (ethnicity and gender)	Experimental
López-López et al. [19]	*A priori* (ethnicity)	Experimental
Muhammad et al. [20]	*A priori* (ethnicity)	Experimental

4. Discussion and Conclusions

In this review, we found that there is a predominant way of approaching the problem of identifying social biases in face recognition systems. This approach is both *aprioristic* and experimental and here we will call it the common approach. In contrast, studies that do not start from assumptions about users' opinions and effects on face recognition (*a posteriori*) [8] and determine their effects in practical cases [8], [9] constitute a minority of cases. As we have seen, one of these studies identified an affected social segment not considered by the *a priori* and experimental studies: it was the workers who "demand significant investments of money, time, and resourcefulness" to “best repair facial verification technology computational failures and errors, and in doing so make themselves machine-readable" [8].

We postulate that this common approach represents a methodological bias that affects the realistic recognition of the problem of social biases in this AI application. This implies that the number of biases may be much higher than that commonly recognized, which boils down to discrimination by gender, age, skin color, or ethnicity. It should be remembered that the common methodology starts from the basis of discrimination against groups recognized by society (women, dark-skinned people, and the elderly) (Table 1); however, there is no reason to believe that there are no other deep-rooted unconscious biases that have never been discussed by society, since this depends on the historical context. This is especially important considering the social categories of gender, race, and age are recognized as "the big three," or the three particularly prominent social categories into which people automatically categorize individuals, although there are infinite ways in which humans can create group distinctions [21][22]. People are actually multidimensional (someone may be a white male, which would make him subject to fewer errors in IA, but an old worker, which would make him more prone to these errors), so the universe of systematic errors may be difficult to describe and mitigate. One could argue that Watkins' study [8] implies that AI errors can actually affect all people, as she found that common characteristics such as hairstyle or glasses can affect the outcome of AI.

We think that users of xAI tools should take this into account, especially since the automation of xAI-derived explanations brings about human overreliance and causes humans to bypass their own correct answers and validate incorrect answers from AI [5], [23]. Furthermore, the cognitive effort to understand certain AI explanations negatively affects the interpretation of recommendations [24]. Thus, explaining why an AI arrives at a result does not ensure that the user comprehends the result. However, it has been shown that users who engage analytically significantly increase the effectiveness of explainable AI [25]. We believe that IA explanations have tremendous potential to facilitate awareness of deep-rooted biases and that this is possible as long as there are conscious users who start with as few assumptions as possible. Here, we postulate that to understand the real dimension of social biases and their effects on AI applications, explainable IA and individuals who are cognitively involved in searching for social biases must work with an *a posteriori* approach and research on real cases. To the best of our knowledge, we are the first to postulate that an *a posteriori* approach can help reveal deep-rooted biases.

Acknowledgements

This study was funded by the ANID - Fondecyt postdoctoral - 3210026.

References

[1] T. Gwyn, K. Roy, Examining Gender Bias of Convolutional Neural Networks via Facial Recognition, Future Internet 14(2022), 375. https://doi.org/10.3390/fi14120375

[2] Georgopoulos, Markos, James Oldfield, Mihalis A. Nicolaou, Yannis Panagakis and Maja Pantic. “Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer”. International Journal of Computer Vision 129(2021): 2288–2307.

[3] K. Baum, S. Mantel, E. Schmidt, T. Speith, From Responsibility to Reason-Giving Explainable Artificial Intelligence, Philosophy and Technology 35(2022), 1–30. https://doi.org/10.1007/s13347-022-00510-w

[4] Gunning, David, Mark Choi Stefik, Jaesik Miller, Timothy Stumpf, Simone Yang and Guang Zhong “XAI-Explainable artificial intelligence”. Science Robotics 4(2019): 4–6.

[5] A. Bertrand, R. Belloum, J. R. Eagan, W. Maxwell, How cognitive biases affect XAI-Assisted decision-making: A systematic review, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, 2022, pp. 78–91, vol. 1. doi: https://doi.org/10.1145/3514094.3534164

[6] Tricco, Andrea C., Lillie Erin Zarin, Wasifa O'Brien, Kelly K. Colquhoun, Heather Levac, Danielle Moher, David Peters, Micah D.J. Horsley, Tanya Weeks, Laura Hempel, Susanne Akl, Elie A. Chang, Christine McGowan, Jessie Stewart, Lesley Hartling, Lisa Aldcroft, Adrian Wilson, Michael G. Garritty, Chantelle Lewin, Simon Godfrey, Christina M. MacDonald, Marilyn T. Langlois, Etienne V. Soares-Weiser, Karla Moriarty, Jo Clifford, Tammy Tunçalp and Sharon E Özge Straus. “PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation”. Annals of Internal Medicine 169(2018): 467–473.

[7] J. Brownlee, Deep Learning for Computer Vision Image Classification, Object Detection, and Face Recognition in Python. Edition v1.3. Machine Learning Mastery, online, 2019.

[8] E. A. Watkins, Face Work: A Human-Centered Investigation into Facial Verification in Gig Work, Proceedings of the ACM on Human-Computer Interaction 7(2023), 1–24. doi: https://doi.org/10.1145/3579485

[9] T.L. Johnson, N.N. Johnson, D. McCurdy, M.S. Olajide, Facial recognition systems in policing and racial disparities in arrests, Government Information Quarterly 39(2022), 101753. doi: https://doi.org/10.1016/j.giq.2022.101753

[10] Albiero, Vitor, Kai Zhang, Michael C. King and Kevin W. Bowyer. “Gendered Differences in Face Recognition Accuracy Explained by Hairstyles, Makeup, and Facial Morphology”. IEEE Transactions On Information Forensics and Security 17(2022): 127–137.

[11] Franco, Danilo, Nicolò Navarin, Michele Donini, Davide Anguita and Luca Oneto. “Deep fair models for complex data: Graphs labeling and explainable face recognition.” Neurocomputing 470(2022): 318–334.

[12] M. Georgopoulos, Y. Panagakis, M. Pantic, Investigating bias in deep face analysis: The KANFace dataset and empirical study. Image and Vision Computing 102(2020), 103954. doi https://doi.org/10.48550/arXiv.2005.07302

[13] D. Celis, M. Rao, Learning facial recognition biases through VAE latent representations, in: Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia, Association for Computing Machinery, New York, NY, 2019, pp. 26–32. doi: https://doi.org/10.1145/3347447.3356752

[14] J. Coe, M. Atay, Evaluating impact of race in facial recognition across machine learning and deep learning algorithms, Computers, 10(2021) 113. Doi: https://doi.org/10.3390/computers10090113

[15] T.P. Pagano, R.B. Loureiro, F.V.N. Lisboa, G.O.R. Cruz, R.M. Peixoto, G.A.D.Guimaraes, E.L. S. Oliveira, I. Winkler, E. G. S. Nascimento, Context-Based Patterns in Machine Learning Bias and Fairness Metrics: A Sensitive Attributes-Based Approach, Big Data And Cognitive Computing 7(2023), 27. doi: https://doi.org/10.3390/bdcc7010027

[16] Wang, Mei, Yaobin Zhang and Weihong Deng. “Meta Balanced Network for Fair Face Recognition”. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(2022): 8433–8448.

[17] I. Serna, A. Morales, J. Fierrez, N. Obradovich. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning, Artificial Intelligence 305(2022) 103682. doi: https://doi.org/10.1016/j.artint.2022.103682

[18] Jiang, Luo, Juyong Zhang and Bailin Deng. “Robust RGB-D Face Recognition Using Attribute-Aware Loss”. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(2020), 2552–2566.

[19] López-López, Eric, Xosé M. Pardo, Carlos V. Regueiro, Roberto Iglesias and Fernando E. Casado. “Dataset bias exposed in face verification”. IET Biometrics 8(2019): 249–258.

[20] Muhammad, Jawad, Yunlong Wang, Caiyong Wang, Kunbo Zhang and Zhenan Sun. “CASIA-Face-Africa: A Large-Scale African Face Image Database”. IEEE Transactions on Information Forensics and Security 16 (2021): 3634–3646.

[21] Taylor, Shelley E., Susan T. Fiske, Nancy L. Etcoff and Audrey J. Ruderman. “Categorical and contextual bases of person memory and stereotyping”. Journal of Personality and Social Psychology 36(1978): 778–793.

[22] Stangor, Charles, Laure Lynch, Changming Duan and Beth Glass. “Categorization of Individuals on the Basis of Multiple Social Features”. Journal of Personality and Social Psychology 62(1992): 207–218.

[23] M. Schemmer, N. Kühl, C. Benz, and G. Satzger. ON THE INFLUENCE OF EXPLAINABLE AI ON AUTOMATION BIAS. arXiv preprint 2204.08859, 2022.

[24] L.V. Herm. IMPACT OF EXPLAINABLE AI ON COGNITIVE LOAD: INSIGHTS FROM AN EMPIRICAL STUDY. arXiv preprints 2304.08861, 2023.

[25] Z. Buçinca, M.B. Malaya, K. Z. Gajos, To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making, Proceedings of the ACM on Human-Computer Interactionvol 5(2021) CSCW1. doi: https://doi.org/10.48550/arXiv.2102.09692