Unveiling the blindspots: Examining availability and usage of protected attributes in fairness datasets*

Jan Simson[1,2], Alessandro Fabris[3] and Christoph Kern[1,2]

[1]LMU Munich, Ludwigstr. 33, 80809 München, Germany
[2]Munich Center for Machine Learning (MCML), Oettingenstraße 67, 80538 München, Germany
[3]Max Planck Institute for Security and Privacy, Universitätsstraße 140, 44799 Bochum, Germany

Abstract

This work examines the representation of protected attributes across tabular datasets used in algorithmic fairness research. Drawing from international human rights and anti-discrimination laws, we compile a set of protected attributes and investigate both their availability and usage in the literature. Our analysis reveals a significant underrepresentation of certain attributes in datasets that is exacerbated by a strong focus on race and sex in dataset usage. We identify a geographical bias towards the Global North, particularly North America, potentially limiting the applicability of fairness detection and mitigation strategies in less-represented regions. The study exposes critical blindspots in fairness research, highlighting the need for a more inclusive and representative approach to data collection and usage in the field. We propose a shift away from a narrow focus on a small number of datasets and advocate for initiatives aimed at sourcing more diverse and representative data.

Keywords

critical data studies, dataset usage, protected groups, generalization

1. Introduction

Algorithmic fairness has become a significant area of research in recent years, with a growing body of work aimed at addressing bias and discrimination in machine learning systems. Identifying and mitigating harmful practices against vulnerable individuals and groups in prediction algorithms lies at the core of this field and to study these issues adequate and nuanced data sources are needed.

In this work, we examine datasets and how they are used within the fairness literature. We present an overview of attributes which are protected by anti-discrimination legislation across multiple continents and study their availability in datasets and usage in fairness research. We identify issues regarding the diversity of protected attributes represented in datasets and their geographic representativeness, highlighting how populations are neglected in the literature.

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

*This is an extended abstract of [1], published at FAccT 2024.

jan.simson@lmu.de (J. Simson); alessandro.fabris@mpi-sp.org (A. Fabris); christoph.kern@lmu.de (C. Kern)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. Methodology

For this work, we collected and manually annotated usage of tabular datasets in fair classification tasks. We built on top of a comprehensive survey of fairness datasets by Fabris et al. [2], leveraging the same inclusion criteria. We focus on tabular datasets used for fair classification in this work, due to their important role in the fairness literature [2, 3]. We study the use of tabular datasets ($N = 36$) across 142 articles. Since datasets appear in multiple publications and most publications use multiple datasets, the total number of dataset and publication combinations examined was $N = 280$, with $n = 233$ instances of sufficient information to reconstruct (or reasonably guess) protected attribute usage.

To define protected attributes, we draw from domain-specific legislation and human rights law. We define as *protected* all attributes which are explicitly mentioned as prohibited drivers of discrimination and inequality. For example, Article 21 of the Charter of Fundamental Rights of the European Union states "Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited" [4].

We try to adress the *Global North* and especially *U.S.* focus in AI ethics and fairness research [5, 6, 7] by including works from the European Union (Charter of Fundamental Rights of the European Union [4], EU legislation on fair hiring [8]), from other continents (African Charter on Human and Peoples' Rights [9], the Arab Charter on Human Rights [10], the ASEAN Declaration of Human Rights [11]) and global works (Universal Declaration of Human Rights [12]) besides works from North America (the American Declaration of the Rights and Duties of Man [13], US fair lending legislation [14]). However, we acknowledge that we fail to mitigate the *Global North* bias completely, given the strong presence of said regions in research.

Drawing from this literature, we provide a shallow categorization of protected attributes, identifying seven main categories (Table 1). It is worth noting that this is not a complete categorization of all protected attributes around the globe and across sectors. Rather, our categorization aims to guide an inclusive discussion of algorithmic fairness research through the lens of protected attributes.

3. Results

The geographical provenance of datasets used in the examined literature is clearly skewed. Among datasets from a single continent, 21 come from North America, 5 from Europe, 2 from Asia, and 2 from South America, confirming a *Global North* and especially North American dominance in AI ethics research [5, 6, 7]. Since fairness is highly contextual, there is a risk that the fairness detection and mitigation strategies built by this research community will not transfer and, therefore, underserve neglected geographical areas [15].

We further notice a highly uneven distribution of both the availability and usage of protected attributes. The left bar chart in Figure 1 depicts protected attributes available in fairness datasets and the right chart their usage in the examined literature. There is a particular focus in both availability (n=17) and usage (n=167) on race as a protected attribute. On the other hand,

Table 1
Protected attributes under global anti-discrimination legislation. Attributes considered protected under international human rights works and anti-discrimination law. We report a tick (✓) when the literal phrasing (in the original law or official clarifications) matches the row header and report a tilde (∼) if a similar concept is present, but wording is different.

	UN Charter [12]	**African Charter [9]**	**Arab Charter [10]**	**ASEAN Declaration [11]**	**American Declaration [13]**	**US Fair Lending [14]**	**EU Charter [4]**	**EU Fair Hiring [8]**
Gender and Sexual Identity								
Sex	✓	✓	✓		✓	✓	✓	✓
Sexual orientation						✓	✓	✓
Gender				✓			∼	∼
Racial and Ethnic Origin								
Race	✓	✓	✓	✓	✓	✓	✓	∼
Color	✓	✓	✓			✓	✓	
Ethnic origin	∼	∼					✓	✓
National origin	✓	✓	✓	✓	∼		∼	
Language	✓	✓	✓	✓	✓	✓		
National minority							✓	
Socioeconomic Status								
Social origin	✓	✓	✓	✓			✓	
Property	✓	∼	∼	∼			✓	
Recipient of public assistance						✓		
Religion, Belief and Opinion								
Religion	✓	✓	∼	✓	∼	✓	∼	∼
Political opinion	✓	✓		✓			✓	
Other opinion	✓	✓	∼	✓			✓	
Family								
Birth	∼	∼	✓	✓			✓	
Familial status							✓	
Marital status							✓	
Disability and Health Conditions								
Disability			✓	✓		✓	✓	✓
Genetic features							✓	
Age								
Age				✓		✓	✓	✓

attributes about *religion, belief and opinion* are entirely missing on both sides. Information on *disability and health conditions* is also infrequently available ($n = 3$) and never used in the surveyed literature. *Socioeconomic status* descriptors are more commonly available yet often neglected. This threatens the applicability of research findings across contexts, as information on race for example is hardly available in EU data [16].[1]

4. Discussion

We unveil blindspots in fairness research, demonstrating a neglect of vulnerable subpopulations in the literature. We will further present additional results from our data collection at the conference, indicating other troubling practices in the field, such as a lack of sufficient reporting

[1] There was also a small number of protected attributes used in the literature but not referenced in legislation, such as employment status, alcohol consumption, neighborhood, body-mass index, and profession.

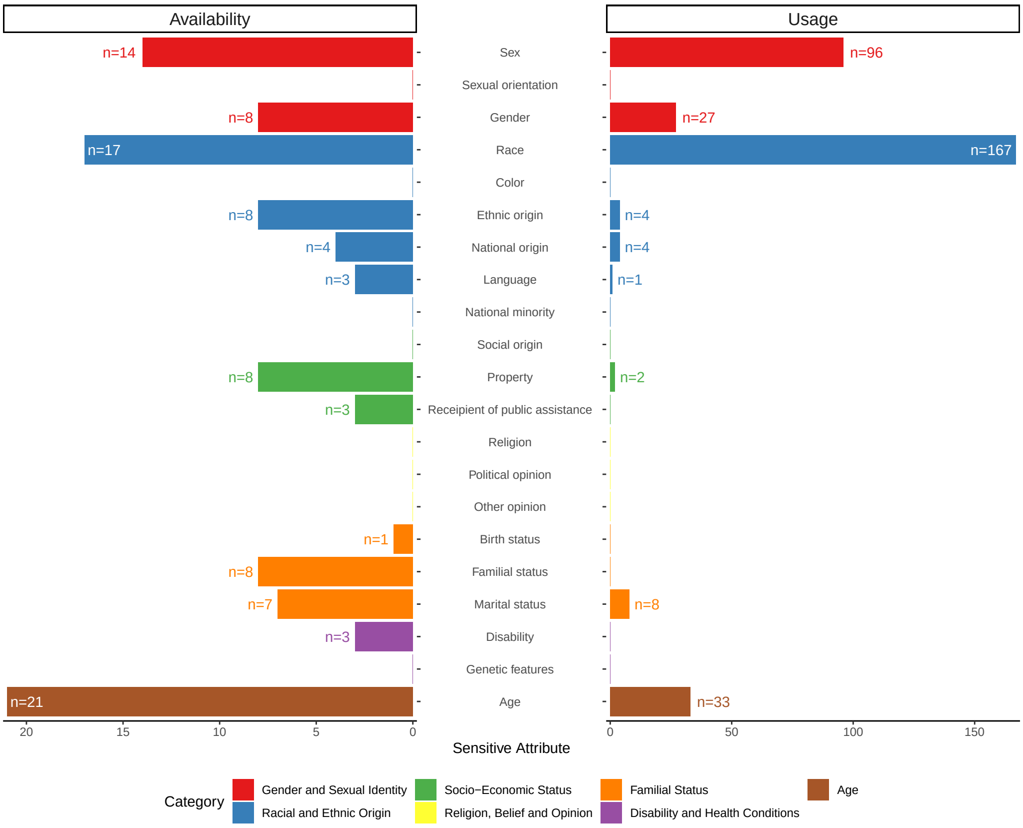

Figure 1: There is a stark difference between attributes considered protected under international legislation and their availability, as well as usage in datasets. Bar charts displaying the availability in datasets (left) and usage in the literature (right) of protected attributes for all categories of protected attributes in Table 1.

of dataset usage impacting reproducibility and potentially harmful practices in the processing of protected attributes leading to a neglect of minorities.

While valid reasons exist against the collection of protected data [17], motivating e.g. the line of work on fairness under unawareness [14, 18], we believe they are not sufficient to explain the observed lack in usage of particular attributes. We observe a clear trend towards certain protected attributes being more readily available in datasets which, however, is amplified by a strong tendency of papers to (1) repeatedly focus on the same small number of datasets and (2) especially rely on race and sex as protected attributes. It is worth noting that this trend extends to fairness research more broadly, including qualitative studies. These practices also have a tendency to self-reinforce, increasing the likelihood of future research to conform. Recent articles published at fairness conferences, such as *FAccT* (the ACM Conference on Fairness, Accountability, and Transparency) and *AIES* (the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society), for example, mention race and gender by an order of magnitude more frequently than religion, disability, socioeconomic status, and sexual orientation [19].

We argue for a move towards a research roadmap to tackle these issues within the complex social, legal and technical landscape they reside in (as advocated, for example, in Guo et al. [20]). In particular, we propose a move away from focusing exclusively on a small number of datasets[2], such as Adult, German Credit or COMPAS. Instead we suggest an increased focus on using a diverse set of datasets and sourcing more representative data to fill in the gaps of available datasets. We call for dedicated initiatives, including for example data donation campaigns and citizen science initiatives, capable of filling this gap and responsibly handling the collected data. We refer readers to the full paper [1] for a more nuanced discussion. A list of datasets and their protected attributes, as well as further analyses are available on Github.

References

[1] J. Simson, A. Fabris, C. Kern, Lazy data practices harm fairness research, in: FAccT '24: 2021 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brasil, June 3-6, 2024, ACM, 2024. URL: https://doi.org/10.1145/3630106.3658931. doi:10.1145/3630106.3658931.

[2] A. Fabris, S. Messina, G. Silvello, G. A. Susto, Algorithmic fairness datasets: the story so far, Data Min. Knowl. Discov. 36 (2022) 2074–2152. URL: https://doi.org/10.1007/s10618-022-00854-z. doi:10.1007/S10618-022-00854-Z.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2022) 115:1–115:35. URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.

[4] Euopeam Union, Charter of fundamental rights of the european union c-364/01, 2000. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32000X1218%2801%29.

[5] C. T. Okolo, N. Dell, A. Vashistha, Making ai explainable in the global south: A systematic review, in: ACM SIGCAS/SIGCHI Conf. on Computing and Sustainable Societies (COMPASS), 2022, pp. 439–452.

[6] C. Roche, D. Lewis, P. Wall, Artificial intelligence ethics: An inclusive global discourse?, arXiv preprint arXiv:2108.09959 (2021).

[7] A. A. Septiandri, M. Constantinides, M. Tahaei, D. Quercia, WEIRD faccts: How western, educated, industrialized, rich, and democratic is facct?, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023, ACM, 2023, pp. 160–171. URL: https://doi.org/10.1145/3593013.3593985. doi:10.1145/3593013.3593985.

[8] A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Z. Borgesius, A. J. Biega, Fairness and bias in algorithmic hiring: a multidisciplinary survey (2024).

[9] Organisation of African Unity, African charter on human and peoples' rights, 1981. https://au.int/sites/default/files/treaties/36390-treaty-0011_-_african_charter_on_human_and_peoples_rights_e.pdf.

[10] Council of the League of Arab States, Arab charter on human rights, 2004.

[11] Association of Southeast Asian Nations, Asean declaration of human rights, 2012. https://asean.org/asean-human-rights-declaration/.

[12] United Nations, Universal declaration of human rights, 1948. https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[13] Organization of American States, American declaration of the rights and duties of man, 1948. https://www.oas.org/en/iachr/mandate/Basics/american-declaration-rights-duties-of-man.pdf.

[14] J. Chen, N. Kallus, X. Mao, G. Svacha, M. Udell, Fairness under unawareness: Assessing disparity when protected class is unobserved, in: danah boyd, J. H. Morgenstern (Eds.), Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019, ACM, 2019, pp. 339–348. URL: https://doi.org/10.1145/3287560.3287594. doi:10.1145/3287560.3287594.

[15] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, V. Prabhakaran, Re-imagining algorithmic fairness in india and beyond, in: M. C. Elish, W. Isaac, R. S. Zemel (Eds.), FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, ACM, 2021, pp. 315–328. URL: https://doi.org/10.1145/3442188.3445896. doi:10.1145/3442188.3445896.

[16] S. Jaime, C. Kern, Ethnic classifications in algorithmic fairness: Concepts, measures and implications in practice, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 237–253. URL: https://doi.org/10.1145/3630106.3658902. doi:10.1145/3630106.3658902.

[17] M. Andrus, E. Spitzer, J. Brown, A. Xiang, What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness, in: M. C. Elish, W. Isaac, R. S. Zemel (Eds.), FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, ACM, 2021, pp. 249–260. URL: https://doi.org/10.1145/3442188.3445888. doi:10.1145/3442188.3445888.

[18] A. Fabris, A. Esuli, A. Moreo, F. Sebastiani, Measuring fairness under unawareness of sensitive attributes: A quantification-based approach, J. Artif. Intell. Res. 76 (2023) 1117–1180. URL: https://doi.org/10.1613/jair.1.14033. doi:10.1613/JAIR.1.14033.

[19] A. Birhane, E. Ruane, T. Laurent, M. S. Brown, J. Flowers, A. Ventresque, C. L. Dancy, The forgotten margins of AI ethics, in: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022, ACM, 2022, pp. 948–958. URL: https://doi.org/10.1145/3531146.3533157. doi:10.1145/3531146.3533157.

[20] A. Guo, E. Kamar, J. W. Vaughan, H. M. Wallach, M. R. Morris, Toward fairness in AI for people with disabilities sbg@a research roadmap, ACM SIGACCESS Access. Comput. 125 (2020) 2. URL: https://doi.org/10.1145/3386296.3386298. doi:10.1145/3386296.3386298.