Ethnic Classifications in Algorithmic Decision-Making Processes

Sofia Jaime[1], Christoph Kern[1]

Ludwig Maximilian University of Munich[1]

Abstract

In this study, we address the challenges and implications of ensuring fairness in algorithmic decision-making (ADM) practices related to ethnicity. We provide an overview of ethnic classification schemes in European countries and emphasize how the distinct approaches to ethnicity and race in Europe can impact fairness assessments in ADM. Using German data, we train machine learning classifiers and explore the fairness implications of different ethnic classifications in labor market- and health-related prediction tasks using common group-based fairness metrics. The findings contribute to the understanding of fairness in ADM from a European perspective.

Keywords

ethnicity, algorithmic decision-making, fairness, Europe

Introduction

The conceptualization, measurement and use of protected attributes is at the center point of ethical and legal concerns that have been raised in the context of algorithmic decision-making (ADM). One of the most contentious debates among computer scientists has been ignited by the controversies surrounding the use of (correlates of) ethnicity in machine learning models and its potential implications for fairness in ADM processes. As prominent ADM applications – such as the COMPAS case [1] – originated in the U.S., these discussions typically center around biases towards groups that are defined by *racial categories*. The protection of groups based on race follows U.S. legislation (e.g., the Fair Housing Act or Equal Credit Opportunity Act [2]) and is reflected in the common inclusion of racial information in national surveys and other U.S. data products. Accordingly, previous social-scientific perspectives on ethnic biases in ADM focused on the conceptualization of race and its implications in the U.S. context [3, 4, 5].

Nonetheless, ethnicity-related attributes have been considered when developing ADM in Europe. For example, in the Netherlands, the System Risk Indication (SyRI) was a data analytics system developed by the Dutch government to detect potential welfare fraud and other irregularities [6]. The inclusion of ethnicity as a data point raised concerns about potential discrimination and profiling. Critics argued that using ethnicity as a factor in the analysis could lead to unfair targeting and stigmatization of specific ethnic groups and that SyRI violated privacy rights and disproportionately targeted vulnerable communities, leading to stigmatization and discrimination.

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

sofia.jaime@stat.uni-muenchen.de (S. Jaime); christoph.kern@stat.uni-muenchen.de (C. Kern)

Against this background, it is important to note that Europe has a fundamentally different approach to ethnicity and race than the U.S., which may lead to different issues and fairness challenges in algorithmic decision-making. These differences do not only relate to legal frameworks, but also affect the operationalization of ethnicity in modeling practices as information regarding the race of a person is hardly included in data sets that have been collected in European contexts.

First, the European Union has implemented a data protection law, the General Data Protection Regulation (GDPR), whose goal is to safeguard individuals' personal data. It aims to harmonize data protection laws, empower individuals, and impose obligations on organizations handling personal data. Among the principles of GDPR, one key principle is the requirement for lawful, fair, and transparent data processing [7]. This means that organizations must process personal data in a legal and ethical manner, ensuring individuals are informed about how their data is being used. Another crucial principle is the need to ensure the integrity and confidentiality of personal data, which means that organizations must implement appropriate security measures to protect against unauthorized access, loss, or destruction of data. Article 9 of the GDPR outlines the definition of sensitive data as personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a person, data concerning health, or data concerning a person's sex life or sexual orientation. At the national level, countries such as Germany have an anti-discrimination legislation, which does refer to both race and *ethnic origin* as protected attributes [8].

Second, in Europe, collecting data specifically on race can be complex and sensitive [9, 10]. Historical reasons have influenced the limited collection of race-related data, such as the colonial legacy of many European countries has made discussions around race complex and sensitive, leading to hesitations in collecting data on race. Another fact is the post-World War II focus on moving beyond racial divisions and fostering inclusive societies. In the aftermath of World War II, European countries placed emphasis on rebuilding and promoting principles of equality, non-discrimination, and human rights.

These historical reasons, encompassing colonial history and post-war focus, as well as legal and ethical considerations, conceptual challenges, and data protection laws, have collectively contributed to the limited collection of race-related data in Europe. Therefore, it remains unclear how the concept of protecting racial minorities should be translated to (and implemented in) fair machine learning applications in a European setting. Studying this mapping is particularly important as the utilization of different classifications can impact fairness metrics. Concretely, existing biases may be obscured dependent on the exact definition of the protected groups and the measures that are used for implementing group classifications.

We fill this gap in the literature by providing a comprehensive overview of ethnic classification schemes in European contexts and presenting an empirical use case that examines fairness in ADM across ethnic classifications with German data. Our study delves into the fairness implications of utilizing different ethnic classifications in automated decision-making processes. We aim to understand how these classifications can affect the (apparent) fairness of predictions made by algorithmic systems, and thus the susceptibility of fairness metrics to different operationalizations of ethnicity. With this research, our goal is to add to the knowledge on the difficulties and intricacies of ensuring fairness in algorithmic decision-making with a focus on ethnic classifications in non U.S. contexts.

Methods

We understand ethnicity as a multi-faceted concept which is manifested through different indicators [11]. We consider that there is no one single cultural criterion which is enough to define an ethnicity [12]. Instead, we regard ethnicity as a complex concept compounded by a number of different domains [13]. Dimensions of ethnicity may include race (or color or visibility), national identity [14], parentage or ancestry [15], nationality, citizenship [16, 17], religion, language [18, 19], and country of birth (or being an immigrant), as well as culture [13]. In this rationale, all the dimensions of ethnicity share a common underlying root, ancestry or origins or "community of descent" [11, 20].

In practice, the absence of agreement among social scientists on how ethnicity should be conceptualized has resulted in varying methods of measurement across European countries. In the UK, the term "ethnic groups" and "ethnic identity" are more widely used, while in Germany, "migration-background" is the most commonly used approach. These various ethnic classification schemes draw on (combinations of) information regarding country of birth, citizenship, nationality of the individuals and their parents. In a fair ADM context, each classification induces its own way of defining protected groups based on the broader concept of ethnicity. We set out to provide a systematical overview of these classifications, their implementations in practice and discuss potential fairness implications that come with different conceptualizations of ethnicity.

On this basis, we present an empirical use case to study the consequences of using different ethnic classifications in fair ADM practice. We take inspiration from the prominent UCI Adult data and its successors [21] and set up two labor market-related prediction tasks, i.e. income and unemployment classification, and a health-related prediction task using German survey data. We next implement a set of ethnic classifications to define protected groups that draw on different measures of ethnic origin (e.g. migration background, nationality, citizenship). We train machine learning classifiers for both prediction tasks and consider common group-based fairness metrics (e.g., statistical parity, equal opportunity difference) for our fairness evaluation. This setup allows us to study the strictness of and overlap between ethnic classifications empirically, and, most importantly, their effect on assessing bias and fairness of the prediction models when defining protected groups based on different classifications.

In summary, our work studies fairness in ADM explicitly from a European perspective which is particularly lacking in current debates on the role of ethnicity in fairness audits of machine learning models. We present different notions of ethnic origin, their practical implementations as well as fairness implications of the use of different classification schemes in practice.

References

[1] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, Ethics of Data and Analytics (2016) 254–264.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.

[3] I. F. Ogbonnaya-Ogburu, A. D. Smith, A. To, K. Toyama, Critical Race Theory for HCI,

Conference on Human Factors in Computing Systems - Proceedings (2020) 1–16. doi:10.1145/3313831.3376392.
[4] S. Benthall, B. D. Haynes, Racial categories in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, 2019, pp. 289–298. URL: http://arxiv.org/abs/1811.11668http://dx.doi.org/10.1145/3287560.3287575https://dl.acm.org/doi/10.1145/3287560.3287575. doi:10.1145/3287560.3287575.
[5] A. Hanna, E. Denton, A. Smart, J. Smith-Loud, Towards a critical race methodology in algorithmic fairness, FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020) 501–512. doi:10.1145/3351095.3372826.
[6] M. van Bekkum, F. Z. Borgesius, Digital welfare fraud detection and the Dutch SyRI judgment, European Journal of Social Security 23 (2021) 323–340. doi:10.1177/13882627211031257.
[7] E. Union, General data protection regulation (2016). URL: https://gdpr-info.eu/art-9-gdpr/.
[8] C. Orwat, Diskriminierungsrisiken durch verwendung von algorithmen, Eine Studie, erstellt mit einer Zuwendung der Antidiskriminierungsstelle des Bundes (2019).
[9] P. Simon, The failure of the importation of ethno-racial statistics in europe: debates and controversies, Ethnic and Racial Studies 40 (2017) 2326–2332.
[10] P. Simon, Collecting ethnic statistics in europe: a review, Ethnic and Racial Studies 35 (2012) 1366–1391.
[11] A. Morning, Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round, Population Research and Policy Review 27 (2008) 239–272. URL: http://link.springer.com/10.1007/s11113-007-9062-5. doi:10.1007/s11113-007-9062-5.
[12] S. L. Schneider, A. F. Heath, Ethnic and cultural diversity in Europe: validating measures of ethnic and cultural background, Journal of Ethnic and Migration Studies 46 (2020) 533–552. URL: https://doi.org/10.1080/1369183X.2018.1550150. doi:10.1080/1369183X.2018.1550150.
[13] J. Burton, A. Nandi, L. Platt, Measuring ethnicity: challenges and opportunities for survey research, Ethnic and Racial Studies 33 (2010) 1332–1349. URL: https://www.tandfonline.com/doi/full/10.1080/01419870903527801. doi:10.1080/01419870903527801.
[14] J. W. Jackson, E. R. Smith, Conceptualizing social identity: A new framework and evidence for the impact of different dimensions, Personality and Social Psychology Bulletin 25 (1999) 120–135.
[15] R. Berthoud, Defining ethnic groups: Origin or identify? (1998).
[16] W. Kymlicka, Three forms of group-differentiated citizenship in canada, Democracy and difference: contesting the boundaries of the political 127 (1996).
[17] Y. Hussain, P. Bagguley, Citizenship, ethnicity and identity: British pakistanis after the 2001 'riots', Sociology 39 (2005) 407–425.
[18] J. S. Phinney, I. Romero, M. Nava, D. Huang, The role of language, parents, and peers in ethnic identity among adolescents in immigrant families, Journal of youth and Adolescence 30 (2001) 135–153.
[19] P. Vedder, E. Virta, Language, ethnic identity, and the adaptation of turkish immigrant youth in the netherlands and sweden, International journal of intercultural relations 29 (2005) 317–337.
[20] D. A. Hollinger, National culture and communities of descent, Reviews in American

History 26 (1998) 312–328.
[21] F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning, 2021. URL: https://arxiv.org/abs/2108.04884. doi:10.48550/ARXIV.2108.04884.