Ethnic Classifications in Algorithmic Fairness: Concepts, Measures and Implications in Practice

Sofia Jaime*
University of California
Irvine, USA
sofiajaime.94@gmail.com

Christoph Kern*
LMU Munich
Munich, Germany
Munich Center for Machine Learning (MCML)
Munich, Germany
University of Mannheim
Mannheim, Germany
christoph.kern@lmu.de

ABSTRACT

We address the challenges and implications of ensuring fairness in algorithmic decision-making (ADM) practices related to ethnicity. Expanding beyond the U.S.-centric approach to race, we provide an overview of ethnic classification schemes in European countries and emphasize how the distinct approaches to ethnicity in Europe can impact fairness assessments in ADM. Drawing on large-scale German survey data, we highlight differences in ethnic disadvantage across subpopulations defined by different measures of ethnicity. We build prediction models in the labor market, health, and finance domain and investigate the fairness implications of different ethnic classification schemes across multiple prediction tasks and fairness metrics. Our results show considerable variation in fairness scores across ethnic classifications, where error disparities for the same model can be twice as large when using different operationalizations of ethnicity. We argue that ethnic classifications differ in their ability to identify ethnic disadvantage across ADM domains and advocate for context-sensitive operationalizations of ethnicity and its transparent reporting in fair machine learning (ML) applications.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; • **Computing methodologies → Machine learning**; • **Applied computing → Law, social and behavioral sciences**.

KEYWORDS

Ethnic Classifications, Fairness Evaluation, Protected Attributes, Ethnicity

ACM Reference Format:
Sofia Jaime and Christoph Kern. 2024. Ethnic Classifications in Algorithmic Fairness: Concepts, Measures and Implications in Practice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3630106.3658902

*Both authors contributed equally to the paper.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658902

1 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 has been ignited by the controversies surrounding the use of (correlates of) ethnicity in machine learning (ML) models and its potential implications for fairness in ADM processes. As prominent ADM applications – such as the COMPAS case [3] – originated in the U.S., these discussions typically center around biases towards groups that are defined by racial categories. While U.S. anti-discrimination legislation (e.g., the Fair Housing Act or Equal Credit Opportunity Act [14, 40]) include concepts beyond race (such as national origin and religion), previous perspectives on ethnic biases in ADM focused predominantly on the conceptualization of race and its implications in U.S. contexts [9, 24, 55].

Nonetheless, ethnicity-related attributes have similarly been used and publicly debated in the context of ADM applications in Europe. The Dutch System Risk Indication (SyRI) program faced criticism for targeting low-income, migrant, and ethnic-minority neighborhoods, leading to concerns about profiling and privacy violations [70]. In the same vein, the Dutch childcare benefits scandal involved an algorithm that used citizenship as a risk factor, which has been tied to false fraud allegations [61]. Such cases underscore the importance of ensuring fairness with respect to ethnicity in ADM systems in Europe [72].

The development of non-discrimination law in the European Union, which began with the Rome Treaty in 1957, was the foundation for a legal framework that influences Europe's handling of ethnicity and racial data. This Treaty, along with the Amsterdam Treaty in 1999, expanded protections against discrimination on various grounds including both racial and ethnic origin. The core of EU non-discrimination law today is composed of directives like the Race Equality Directive 2000/43/EC [49], the Framework Equality Directive [48], and the gender equality Directives 2004/113/EC and 2006/54/EC [50]. These laws, along with relevant articles in the Treaty on EU [52] and the Treaty on the Functioning of the EU [53], and the Charter of Fundamental Rights of the EU [51], collectively reinforce the EU's commitment to non-discrimination and equality. At the national level, countries such as Germany have additional

anti-discrimination legislation, which may similarly refer to both race and ethnic origin as protected attributes [57].

Against the background of evolving EU anti-discrimination laws, it is essential to recognize the differences between the U.S. and European countries in the collection and operationalization of ethnicity in data practice. The U.S. census has traditionally distinguished race and ethnicity as separate categories [43], but announced a shift to a single, combined question for the 2030 census, aiming to better capture the multifaceted identities of the population [54]. In contrast, race is typically not measured in European national censuses or other EU data products [66, 68]. The colonial legacy of many European countries, the focus on moving beyond racial divisions in the aftermath of World War II, legal and ethical considerations, conceptual challenges, and data protection laws have jointly contributed to the restricted collection and use of race-related data in Europe. Therefore, adapting the concept of protecting racial minorities to fair ML applications in Europe presents a challenging and nuanced task.

The ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) has extensively addressed race and fairness in ADM, including papers on racial bias and analysis of racial categories [1, 8, 9, 15, 24, 69], racial bias in NLP tools [12, 20], in computer vision [8, 34], in visual and semantic AI [26, 28, 32, 60, 75, 76], in online platforms and social media [6, 31], and in the healthcare sector [47]. Building on this literature, our research aims to extend the understanding of ethnicity and fairness in ADM by focusing on ethnic classifications and their implications in the European context. Studying operationalizations of ethnicity and their use as protected attributes is particularly important as the utilization of different classifications can impact fairness assessments. Concretely, existing biases may be obscured depending on the exact definition of the protected groups and the measures that are used for implementing group classifications.

In this study, we (1) provide a comprehensive overview of ethnic classification schemes in European contexts and (2) present an empirical case study that examines fairness in ADM across different ethnic classifications. Our main research question is: How do different ethnic classifications impact fairness evaluations in European ADM applications? We aim to understand how these classifications can affect the (apparent) fairness of predictions made by algorithmic systems, and thus the susceptibility of fairness evaluations to different operationalizations of ethnicity. Conceptually, we presume that ethnic classifications differ in their ability to identify ethnic disadvantage across ADM domains and argue for context-sensitive operationalizations of ethnicity in fair ML applications. With this research, our goal is to add to the knowledge on the difficulties and intricacies of ensuring fairness in ADM with a focus on ethnic classifications in non-U.S. contexts.

Our empirical case study exemplifies the consequences of using different ethnic classifications in fair ML practice. We set up four prediction tasks that cover different domains, i.e. labor market, health and finance, using German survey data. We implement a set of ethnic classifications to define protected groups based on different operationalizations of ethnic origin (e.g. direct and indirect migration background, nationality, citizenship) and evaluate prediction models for each prediction task and ethnic classification using prominent group-based fairness metrics. 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 ML models. We present different notions of ethnicity, their practical implementations as well as fairness implications of the use of different classification schemes in practice.

2 CONCEPTS AND MEASURES OF ETHNICITY

2.1 From Race to Ethnicity

Both race and ethnicity are socially constructed classifications with no genetic or scientific basis [2, 23, 56]. Despite their different bases—ethnicity in cultural identity and heritage, and race in physical characteristics and power dynamics—they share a common feature of being products of societal perceptions and structures.

Race and ethnicity, though often overlapping, are distinct concepts. Ethnicity refers to groups of people sharing "a common descent, or having a common national or cultural tradition", as defined in the Oxford English Dictionary [58]. It arises from self-identification (asserted) and group identity, based on internal claims made by the group rather than external attributions [16]. In contrast, race is usually externally imposed (assigned) and closely linked with social power hierarchies. It is often perceived and categorized based on physical characteristics, playing a central role in understanding and addressing racism.

In the United States, race is a prevalent and historically entrenched concept, deeply influenced by the legacy of slavery and colonialism. Race has been socially constructed and institutionalized in ways that have profound impacts on individuals' identities and experiences. This history has created significant racial divides, with systemic barriers impacting people of color. For instance, racial groups in the U.S. often experience differing levels of wealth, health care access, and employment opportunities. However, the U.S. is unique in its method of enumerating population by race and in treating race and ethnicity as different types of identity, a practice not commonly found elsewhere [43].

In contrast, in Europe, ethnic and national identities are more commonly addressed due to different historical and social contexts, including periods of colonialism and the impacts of both World Wars. These events have deeply influenced the ways in which European nations perceive and record ethnic and national identities, thus leading to their more prominent feature in census data and public policy [43]. European nations often have an approach to census data focused on ethnic and national identities, capturing the complexity of their populations, which in turn shape policies in areas such as education, social welfare, and immigration.

2.2 Ethnic Classifications in Europe

Ethnicity can be understood as a multidimensional concept [10, 13]. It integrates various aspects such as race (or skin color or visibility), national identity, ancestry, nationality, citizenship, religion, language, and country of birth, as well as culture. This comprehensive view acknowledges the interconnectedness of these factors, underlining ethnicity's complex nature.

In practice, the polysemy and fluidity of terms related to ethnicity result in varied measures across Europe. The term "ethnicity" carries multiple meanings and definitions, varying with context

and interpretation. For example, the U.K. employs detailed self-reported ethnic categories [5], France prohibits ethnic or racial data collection focusing on citizenship [67], the Netherlands focuses on country of birth and parentage [65], while Sweden collects information on citizenship and the birthplaces of individuals and their parents [73]. This multiplicity in the absence of a common conceptualization contributes to the varied methodologies in measuring ethnicity. The outcome is a diverse array of measures, reflecting the complex nature of ethnicity in Europe. In a fair ML context, each measure induces its own way of defining protected groups based on the broader concept of ethnicity.

In Table 1, we present concepts and measures of ethnicity commonly found in Europe. These were selected based on their presence in national censuses, which offer extensive coverage and reflect the practices of national official statistics. Previous research examining census data up to round 2000 shows that 16 out of 37 European countries enumerated ethnicity [43, 44][1] and that there are several dimensions of ethnicity measured in the censuses [27]. This trend continued in the 2010 round, as we found in the European census data compiled by IPUMS [42], with "ethnicity" and "nationality" being key terms. In addition, we looked at the European Social Survey (ESS; 2002–2020), a major European data collection effort, to complement the concepts used in national censuses [19]. It is important to note that we exclude concepts like race and skin color, reflecting their absence in major European databases. Nonetheless, we acknowledge these concepts as part of the broader spectrum that ethnicity encompasses.

Nationality: Nationality refers to the country (or countries) of which a person holds citizenship. In this definition, the main enclave is the geography [46]. From a broader perception, Miller [41] discusses on nationality, by highlighting three different propositions: national identity as a person's "place in the world" (pp. 10); nations as ethical communities with distinct duties owed to fellow-nationals that are more extensive than general human duties; and the political right of national communities to self-determination, ideally through a sovereign state.

Citizenship: The particular legal bond between an individual and their state, acquired by birth or naturalization whether by declaration, choice, marriage or other means according to national legislation [46]. For instance, citizenship can be defined based on ancestry or geography depending on the context. Ancestry refers to nationality acquired by descent or "blood", also known as *jus sanguinis* citizenship, which pertains to the country of legal nationality. Geographically, citizenship can be derived from the place of birth, known as *jus soli* citizenship, relating to the country or area of origin [17].

In the member states of the EU where a distinction is made between citizenship and nationality, citizenship specifically refers to the legal rights and duties of citizens. In countries where a distinction is maintained, "nationality" often denotes a broader sense of belonging, potentially encompassing ethnic or cultural identity, whereas "citizenship" is more narrowly defined in terms of legal status and rights within the state [46].

[1]Bulgaria, Channel Islands (Jersey), Croatia, Estonia, Hungary, Latvia, Lithuania, Luxembourg, Poland, Republic of Moldova, Romania, Russian Federation, Slovenia, Former Yugoslav Republic of Macedonia, Ukraine, United Kingdom, Yugoslavia.

Country of Birth: The ENM Glossary defines it as "the country of residence (in its current borders, if the information is available) of the mother at the time of the birth or, in default, the country (in its current borders, if the information is available) in which the birth took place" [46]. Country of birth is one of the variables which is more common in survey data. Together with the country of birth of mother and father of the respondent, it is usually used to create other variables such as the different generations of migrants.

Ancestry: Ancestry suggests a link to individuals or elements from the past. This is also known as genealogical ancestry in genetics, which considers someone's parents, grandparents, or even great-grandparents [39]. Hence, the question in Table 1 about country of birth of someone's parents can provide some information about their ancestry, but it does not exhaust the nuances of the term. Ancestry encompasses much more than just the birthplace of direct ancestors, it can include family roots, ethnic backgrounds, and historical migrations that might span multiple countries and regions over generations.

Religion: Religion is one indicator of ethnicity given that the meaning communities give their religious beliefs and rituals contribute to affirming their ethnic identities. For instance, reconstructing their traditional places fosters a sense of community, and for immigrants, it can mean a way of adapting to their new environment [35].

Language: In Phinney's et al. [62] conceptualization, language is a contributor of ethnic identity. In this sense, language helps migrants of second and third generation maintain their ethnic language and feed their sense of belonging to their ethnic group.

Culture: The Oxford Learner's Dictionary defines culture as "the customs and beliefs, art, way of life and social organization of a particular country or group" [59]. We conceive culture as the context of ethnicity, which is flexible and can function as a discoursive resource for ethnicities to identify and generate in-group and out-group boundaries.

Derived Measures: The outlined measures of ethnicity are categorical and represent different ways of capturing one specific dimension of this multi-faceted concept. Their application is highly context-dependent. For instance, Switzerland's four official languages illustrate the limitation of language as a cultural indicator of ethnicity, as a French-speaking individual in Switzerland could belong to various ethnic groups. Moreover, two or more dimensions could be combined to generate new measures, which is the case of *migration background*. In Germany, for instance, this categorization combines place of birth of the individuals themselves and their parents as well as an individuals' citizenship [74]. In contrast, in the U.K. the concepts *ethnic identity* or *ethnic group* include aspects of self-identification with an ethnic group and national identity as well as cultural indicators of ethnicity such as religion [2].

The different concepts of ethnicity and their associated measures can be grouped based on their level of strictness in classification and their substantive implications for identifying disadvantage in different social contexts. Specifically, considering solely "functional" measures of ethnicity such as citizenship to construct a binary grouping results in a restrictive categorization that leaves out individuals who may be classified as migrants by the majority population in social processes based on other signals of ethnicity (such as religion and language). Categorizations that additionally

draw on such signals and indicators based on which individuals may be ethnicized result in broader, but also more heterogeneous, classifications, such as migration background.

2.3 Discrimination and Ethnicity

Ethnic discrimination involves treating individuals differently based on their (inferred) ethnic background or heritage. This can include biases, prejudices, stereotypes, or unequal treatment based on cultural, linguistic, or ancestral factors associated with a particular ethnic group. In the ADM context, ethnic discrimination can result in biased data that influences the algorithm's predictions, potentially leading to unjust outcomes in the distribution of interventions or resources. Discrimination can take on various manifestations – *individual, institutional* and *structural* – each illustrating the uneven distribution of power in societies, as individual members of the dominant group engage in discriminatory acts or institutions form policies in detriment of minority groups [63]. However, different manifestations of discrimination do not operate independent of each other. In fact, they can concurrently influence various levels or arenas, exacerbating ethnic disadvantages. The different manifestations of discrimination can be illustrated in the context of the domains we focus on in our empirical case study – the labor market, health and finance sector.

Individual discrimination concerns the actions undertaken by singular individuals or small groups, directed against members of a different group [63]. For instance, consider a situation within the finance sector where a bank employee rejects a loan application from a migrant by resorting to stereotyping based on assumed group averages or on the grounds of prejudice. Here, the bank employee's action, situated within the finance domain, adversely impacts the migrant's loan prospects.

Institutional discrimination entails entrenched biased actions within significant societal institutions. Typically, this type of discrimination emanates from the dominant majority, wielding influence over institutions [63]. In the German context, e.g., certain select groups can opt for either public or private health insurance. Those with higher incomes or specific positions are able to choose private insurance, thereby establishing a connection between social and economic privilege and private healthcare coverage. Generally, the dominant group tends to have higher earnings than minorities, exemplified here by the healthcare system institution that is limiting the possibilities of the minorities.

Structural discrimination refers to patterns of discrimination that are embedded within the social, economic, and political structures of a society [63]. Unlike individual discrimination, which involves specific actions of individuals, structural discrimination encompasses systemic disparities that affect entire groups based on concepts such as ethnicity. In the labor market context, for example, manifold forms of disadvantage such as differences in educational access and institutional practices culminate in systemic differences in economic outcomes and (un)employment based on ethnicity across Europe [25, 30, 36]. Attributes such as income or unemployment histories may in turn be used as proxy variables in allocation decisions (e.g. in lending practices) and thus access to further resources can systematically differ for ethnic minorities in comparison to majority groups.

We hypothesize that ethnic classifications differ in their ability of capturing ethnic disadvantage based on different forms of discrimination across social contexts. Social implications of specific forms of institutional discrimination may be identifiable based on functional indicators of ethnicity such as nationality if, e.g., institutional services are restricted to or tailored towards national citizens. Individual discrimination, however, manifests in social interactions in which various signals of (inferred) ethnicity may form the grounds of differential treatment. In such contexts, drawing on restrictive classifications will leave out individuals who are ethnicized based on socio-cultural cues e.g. related to religion, culture or language.

3 METHODOLOGY

3.1 Case Study: Ethnic Classifications in Germany

We design an empirical case study to investigate the implications of using different ethnic classifications in fair ADM practice using German survey data. Germany's ethnic diversity is shaped by post-World War II migration, influenced by policies like the 1950s Guest Worker Program and the reformed 2000 Law on Nationality, its asylum policies and significant refugee intake during crises, notably the 2015 Syrian conflict [45]. Studying ethnicity in the German context provides insights into the complexities of fairness-aware ML research in multicultural societies.

3.2 Data Source

We use data from the 2018 wave of the Socio-Economic Panel study (SOEP v37), conducted by the German Institute for Economic Research (DIW), as the basis for this study's analysis [22]. SOEP data consists of multiple probability-based samples of individuals residing in Germany, yielding a high-quality data source with good representation of the German population. Specific subsamples were drawn and invited to participate that explicitly target different migrant populations (see subsection below). The 2018 wave of the SOEP collected information on various dimensions, including sociodemographic characteristics, employment status, income, education, health, and subjective well-being. We provide more detail on the SOEP data in Appendix A.

3.3 Outcome Variables, Predictors and Ethnic Classifications

Outcomes. We select four outcomes that cover different domains: the labor market (high income, unemployed), health (private health insurance) and finance (loan payoff) domain. Our set of outcomes is inspired by common prediction examples in the fair ML literature, such as the tasks presented by Ding et al. [18] in their extension of the UCI Adult data set. In our examples, we aim to mimic the prediction step of an ADM system in which ("high risk") individuals are to be identified which are eligible for receiving interventions or resources.

From a modeling perspective, the outcomes represent targets that vary in their predictability given standard sets of predictor variables and also differ in their respective base rates. We set up prediction tasks with the following binary outcome variables:

Table 1: Concepts and measures of ethnicity in European countries. Survey measures sourced from the ESS [19], except for the nationality metric that comes from the 2011 French National Census. Comparable nationality question wording is also present in the 2000 and 2010 census waves in Ireland, Portugal, and Spain [42].

Concept/Dimension	Definition	Survey Measure
Nationality [41]	The country (or countries) of which a person holds citizenship	What is your nationality?
Citizenship [29]	The particular legal bond between an individual and their state, acquired by birth or naturalisation, whether by declaration, choice, marriage or other means according to national legislation	What citizenship do you hold?
Country of birth [46]	Country where birth took place	In which country were you born?
Ancestry [39]	Heritage, family origins	In which country was your father born?
Religion [35]	Religious believes	Do you consider yourself as belonging to any particular religion or denomination?
Language [62]	Communication system (spoken/read/understood)	What language or languages do you speak most often at home?
Culture [59]	Shared beliefs, practices, norms, values, customs, arts, history, and knowledge	Do you belong to a minority ethnic group in [country]?

(1) High income: This outcome measures whether an individuals' gross income exceeds 3,500 Euro (1 = "high income", 0 = "medium or low income"). The income threshold corresponds to the 75th quantile of the income distribution in the SOEP [18]. Non-working individuals are excluded from the analysis for this outcome.
(2) Unemployed: Indicates whether an individual is registered as unemployed at the employment office (1 = "yes", 0 = "no").
(3) Private insurance: Indicates whether an individual has public or private health insurance (1 = "private", 0 = "public"). In Germany, individuals with high earnings, selected occupations or specific employment types can choose between public and private health insurance. Otherwise, individuals need to have public insurance.
(4) Loan payoff: Measures whether the respondent or someone in their household currently pays back loans and interest that were taken for large purchases or other expenditures (1 = "yes", 0 = "no").

Predictors. We use a set of standard socio-demographic characteristics as features in the prediction models for each outcome: Age, sex, years of education, marital status, type of household and state of residence (see also Ding et al. [18]). For the purposes of this study, all of these variables are equally considered as non-sensitive attributes. In addition to this basic set of features, we consider task-specific predictors as outlined in Table 3 in the appendix.

Ethnic Classifications. We use multiple measures of ethnicity that will be used to define protected groups in our fairness evaluations of the prediction models. The resulting ethnic classifications are inspired by the concepts presented in section 2.2, adapted to the information and indicators available in the SOEP data. We construct five binary ethnic classifications:

(1) Direct migration background: This variable indicates whether an individual has directly moved from any country to Germany. It distinguishes individuals with direct migration experience from those who are either the offspring of migrant/s or have no migration history at all (1 = "direct migration background", 0 = "no or indirect migration background").
(2) Direct or indirect migration background: Indicates whether an individual is a migrant and/or is a child of a migrant/s in Germany. The counterpart is the group that has no migration history at all (1 = "direct or indirect migration background", 0 = "no migration background").
(3) Nationality: The classification indicates the legal affiliation of an individual with Germany (1 = "nationality other than German", 0 = "German nationality").
(4) First or second citizenship: It indicates if an individual's first or second citizenship is non-German, relative to the group whose citizenship is German (1 = "(1st or 2nd) citizenship other than German", 0 = "German citizenship").
(5) Migration sample: This variable indicates whether an individual is part of a sample of the SOEP study that particularly targeted a migrant population (1 = "SOEP migration sample", 0 = "other SOEP sample"). Typically, in such samples at least one person of the respondents' household belongs to a migrant population that may be defined by nationality and/or the year of immigration into Germany. Specifically, we consider samples B, D and M1 to M5 as SOEP migration samples.[2] See Kara et al. [33] for an overview of SOEP samples.

In this context, the classifications based on "first or second citizenship" and especially "nationality" may be viewed as functional or restrictive groupings, while the measures "direct or indirect migration background" and "migration sample" are less restrictive as they define ethnicity not only on legal grounds but also by considering family origins and the household context. Summary statistics for the outcome variables, predictors and ethnic classifications are provided in Table 4 and Table 5 in the appendix.

[2]The SOEP labels of these samples are as follows: B: "Foreigners in the Federal Republic of Germany", D: "Immigrants", M1, M2: "Migration Sample", M3, M5: "Refugee Sample", M4: "Refugee Family Sample".

3.4 Prediction Models

We follow the standard supervised learning pipeline to build prediction models for each outcome variable. The SOEP data is split into a training set (75%) used for model tuning and training and a test set (25%) for model performance and fairness evaluation. We use random forests [11] for building prediction models, which are tuned using 5-fold cross-validation in the training set, optimized with respect to ROC-AUC. The respective best model is used to predict the outcome in the test set, where we threshold the scores at the 75% quantile to obtain class predictions. We repeat the model training and evaluation process 10 times with different random train-test splits and report average performance and fairness scores to improve robustness [21]. The various ethnic classifications are not used in model training and enter the modeling pipeline only at the evaluation step.

3.5 Prediction Performance and Fairness Metrics

We evaluate the performance of the prediction models using standard measures: ROC-AUC and balanced accuracy ($BACC := 1/2(TPR+TNR)$). Next to overall performance, we compute subgroup-specific balanced accuracy scores using the different ethnic classifications as an initial assessment of potential fairness concerns.

Our main focus is on the implications of different operationalizations of ethnicity in fairness evaluations. We consider a set of common group-based fairness metrics, and compute each measure using the different ethnic classifications presented above to define the respective protected and unprotected group. As we study multiple outcomes, we consider different fairness concepts and use metrics that follow the independence, separation and sufficiency criteria [7, 38].

In our fairness evaluations, we first take the difference in balanced accuracy scores between the respective protected (s^*) and unprotected group (s) as measured by different classifications ($S(c)$). In this evaluation, we investigate whether different ethnic classifications lead to different assessments of performance gaps across subgroups.

- Balanced Accuracy (BACC) Difference := $BACC(s^*) - BACC(s)$

We consider the independence criterion by computing demographic parity across ethnic classifications. In our setting, different ethnic classifications may imply variation in the average risk predictions across protected groups due to differences in respective subgroup-specific base rates and/or due to differences in the amplification of group differences by the prediction models.

- Demographic Parity := $P(\hat{Y}|S^{(c)} = s^*) - P(\hat{Y}|S^{(c)} = s)$

We next compute separation-based metrics. Differences in these measures between ethnic classifications may occur when a model discounts the fitness of protected groups differently across classification, e.g. due to group-specific label bias and differences in training data quality.

- False Negative Rate (FNR) Difference := $P(\hat{Y} = 0|Y = 1, S^{(c)} = s^*) - P(\hat{Y} = 0|Y = 1, S^{(c)} = s)$
- False Positive Rate (FPR) Difference := $P(\hat{Y} = 1|Y = 0, S^{(c)} = s^*) - P(\hat{Y} = 1|Y = 0, S^{(c)} = s)$
- Equal Opportunity Difference (EOD). Mean of absolute difference in FNR and absolute difference in FPR.

Our last set of fairness metrics considers sufficiency. While differences in these measures across ethnic classifications may occur due to similar reasons as for separation-based metrics, their practical implications may differ depending on whether separation or sufficiency is considered as the main fairness objective.

- Positive Predictive Value (PPV) Difference := $P(Y = 1|\hat{Y} = 1, S^{(c)} = s^*) - P(Y = 1|\hat{Y} = 1, S^{(c)} = s)$
- Negative Predictive Value (NPV) Difference := $P(Y = 0|\hat{Y} = 0, S^{(c)} = s^*) - P(Y = 0|\hat{Y} = 0, S^{(c)} = s)$
- Predictive Parity (Pred. Parity). Mean of absolute difference in PPV and absolute difference in NPV.

Equal opportunity and predictive parity difference are in range $[0, 1]$ and all other fairness metrics in range $[-1, 1]$, with zero representing the (typically favorable) case of no differences between protected and unprotected group in the respective quantity that is being compared.

3.6 Software

We used R (4.3.2) [64] for data preparations, model training and evaluation, using the tidyverse (2.0.0), ranger (0.16.0), mlr3 (0.17.1) and mlr3fairness (0.3.2) packages. Code for replication purposes is available at the following OSF repository: https://osf.io/4wvym/?view_only=6f6e6068850444b98a60ec6b2061089e.

4 RESULTS

4.1 Ethnic Classifications, Intersections and Base Rates

We firstly highlight how different measures of ethnicity capture subpopulations that do not only differ in size and inclusiveness but also in the degree of ethnic disadvantage. In Figure 1, we present base rate plots that show the distribution (relative frequencies) of the four outcome variables by ethnic classification, in reference to the overall distribution of these outcomes across the full sample. Overall, the considerable differences in the share of high income, unemployed and privately insured individuals in any ethnic classification in comparison to the majority population highlight the significant disadvantage of ethnicized people in Germany. Nonetheless, differences between classifications emerge as well. The share of unemployed individuals is lower when the classifications "direct migration background" ($n = 8,154$) and particularly "direct or indirect migration background" ($n = 9,985$) are applied, in comparison to the remaining measures. Classifying individuals based on non-German "nationality" ($n = 6,401$) results in the lowest share of high income and private insurance and the highest unemployment rate. These differences are weakened when the ethnic group is defined just slightly differently by considering individuals with non-German "first or second citizenship" ($n = 7,423$). Defining ethnicity based on "migration sample" membership ($n = 7,259$) (only) stands out with respect to the finance-related outcome as it results in the group with the lowest share of individuals who currently payoff a loan.

The applied ethnic classifications further differ in their level of inclusiveness and mutual overlap. In Figure 3 in the appendix, we

depict the extent of overlap between ethnic groups as measured by their Jaccard similarities, with darker shades indicating a stronger degree of set similarity. Firstly, we observe a pronounced overlap between the "direct or indirect migration background" group and the "direct migration background" group, which is expected as the first group includes the latter by definition. Similarly, the "first or second citizenship" and "nationality" groups exhibit a substantial degree of overlap as the first group extends the latter in our data. However, it also becomes apparent that measures that draw on "nationality", "direct or indirect migration background" and "migration sample" membership do target distinct populations with considerably lower mutual overlap.

Instead of considering each set of ethnic classification separately, the different measures can also be used to study set intersections. In Figure 4 in the appendix we show an intersection plot and associated base rate plots for each outcome variable. The lower part of the plot characterizes the intersections in terms of the underlying combination of ethnicity measures and plots their cardinality (set size). In our data, the largest group is defined by the intersection of all five measures, i.e. includes individuals who are jointly classified as "migrants" by all measures ($n = 5,246$). However, there are also sizable groups which are, e.g., solely classified as migrants based on the "indirect migration background" measure ($n = 891$) or based on all measures except for "migration sample" ($n = 812$). As shown by the base rate plots in the upper part of the figure, these distinct subgroups differ in their distribution of the outcome variables, implying that the way how ethnicity is defined in modeling and evaluation practice has non-trivial implications regarding which specific populations are eventually considered or neglected.

4.2 Subgroup Accuracy and Fairness Evaluation

We next study how fairness evaluations of prediction models depend on the ethnic classification that is used to define protected groups. We start by presenting (subgroup-specific) prediction performance of our four prediction models in Table 2. Overall, our prediction tasks varied in complexity, and we observe high performance scores (ROC-AUC and balanced accuracy) for the high income, unemployed and private insurance outcome, but rather low performance for the model predicting loan payoff. However, examining performance across distinct ethnic classifications uncovers a nuanced narrative. Subgroup-specific performance scores are consistently lower than the respective global score, raising fairness concerns as predictions under any ethnic classification are more often inaccurate. Furthermore, the degree as to which disparate error become visible differs between classifications. Across outcomes, the (relatively) best subgroup-specific performance is recorded based on the classifications "direct or indirect migration background" and "migration sample", while the lowest scores are obtained for the nationality-based grouping. For the loan payoff model, however, the lowest performance is recorded for the "migration sample" for which the model is no better than random guessing.

We next present a visualization of fairness metrics, evaluated under different ethnic classifications, for each of the four models in Figure 2. First, examining the high income model reveals considerable variation in fairness scores across ethnic classifications (Figure 2a). The most extreme (unfair) scores are associated with the "nationality" group, suggesting more pronounced disparities based on this classification. In contrast, the "direct or indirect migration background" and "migration sample" classification show lower unfairness scores. Switching from the "migration sample" to the nationality-based grouping can lead to a doubling in unfairness scores which can imply very different conclusions in fairness auditing practice. The differences in fairness scores are more pronounced than differences in subgroup-specific accuracy, and demonstrate that the different ways of measuring ethnicity can have a profound impact on how a prediction model scores on common metrics in fairness evaluations. For the high income model, more restrictive ethnic classifications tend to reveal stronger error disparities than less restrictive classifications.

Second, the unemployment model generally leads to overly pessimistic predictions for all ethnic groups (Figure 2b). We observe a marked degree of unfairness across all ethnic classifications particularly in metrics like demographic parity and false positive rate difference. However, there is also variation across ethnic classifications, with the most substantial differences appearing between the "nationality" and "direct or indirect migration background" category.

Third, the private insurance model exhibits similarities to the previous models in terms of fairness considerations (Figure 2c). Just as "nationality" played a pivotal role in revealing disparities in the high income and unemployment model, this classification shows the most unfair scores in the private insurance model. It is noteworthy that even a seemingly small change in the definition of the protected ethnic group – from "nationality" to "first or second citizenship" – can significantly affect fairness metrics as it leads to considerably lower false negative rate differences in this case.

Last, the loan payoff model unveils nuanced distinctions when compared to the other three models (Figure 2d). The ethnic group which exhibits the most pronounced disparities is the "migration sample", marking a departure from the pattern observed in the prior models. The fairness scores of this classification may be related to its markedly lower loan payoff base rate compared to the other ethnic groups (Figure 1d). Notable disparities can also be observed based on the "nationality" and "direct migration background" classification.

Figure 5 in the appendix shows fairness metrics computed based on (exclusive) intersections of ethnic classifications. It is highlighted how groups of individuals who are classified as "migrants" by multiple measures jointly experience particularly strong error disparities. However, pronounced unfairness scores are also visible for individuals who are exclusively classified as "migrants" based on migration background measures (fourth column in Figure 5), highlighting the merit and importance of more inclusive classifications next to e.g. "nationality".

5 DISCUSSION

We investigated the complexities of studying algorithmic (un)fairness with respect to ethnicity in European contexts. We argue that studying ethnic disadvantage in ADM applications is fundamentally different to measuring racial biases as typically done in current fair ML practice. In European data sources, race is not measured but instead substituted by a plethora of different measures and indicators of ethnicity with which different ethnic classifications can be formed.

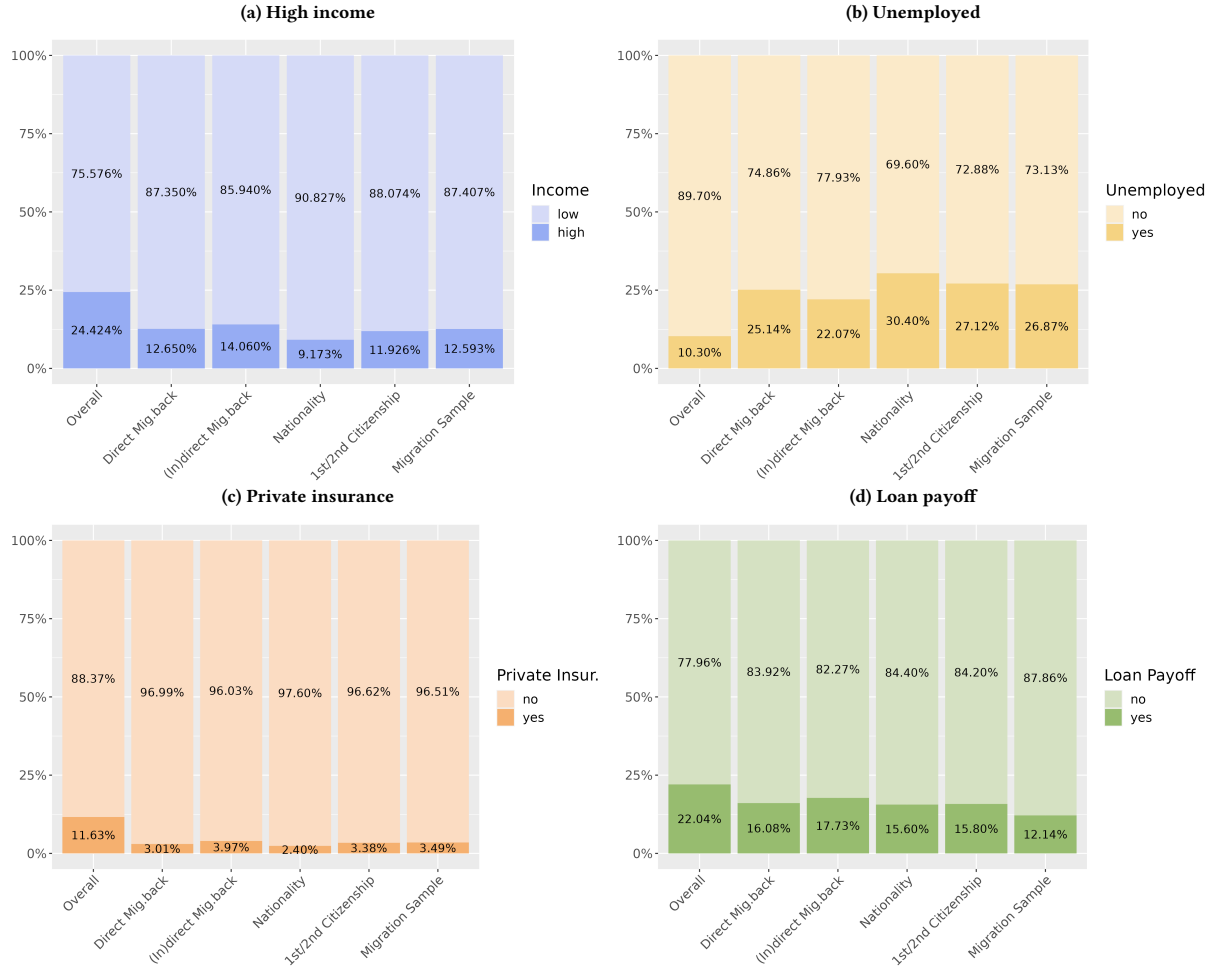

Figure 1: Different ethnic classifications capture different degrees of disadvantage. Plots showing base rates of four outcome variables for the full sample (overall) and by ethnic classification.

Table 2: Variation of prediction performance by ethnic classifications. Performance measures are computed for the full test set (overall) and separately for subgroups defined by different ethnic classifications. Average scores over 10 model runs are reported.

	ROC-AUC	Balanced Accuracy					
	Overall	Overall	Direct migration background	Direct or indirect migration background	Nationality	First or second citizenship	Migration sample
High income	0.891	0.78	0.716	0.740	0.698	0.718	0.746
Unemployed	0.914	0.839	0.734	0.755	0.694	0.723	0.732
Private insurance	0.808	0.724	0.641	0.668	0.600	0.656	0.672
Loan payoff	0.65	0.582	0.540	0.552	0.524	0.533	0.500

In our review of measures commonly used in European surveys, we distinguished between more strict, functional measures (nationality, citizenship) and more inclusive measures that also capture signals based on which individuals may be ethnicized (touching concepts such as ancestry, religion, language and culture). Understanding the range of measures and their implications is critical as we argue that different ethnic classifications are needed to adequately capture ethnic disparities in different ADM application domains:

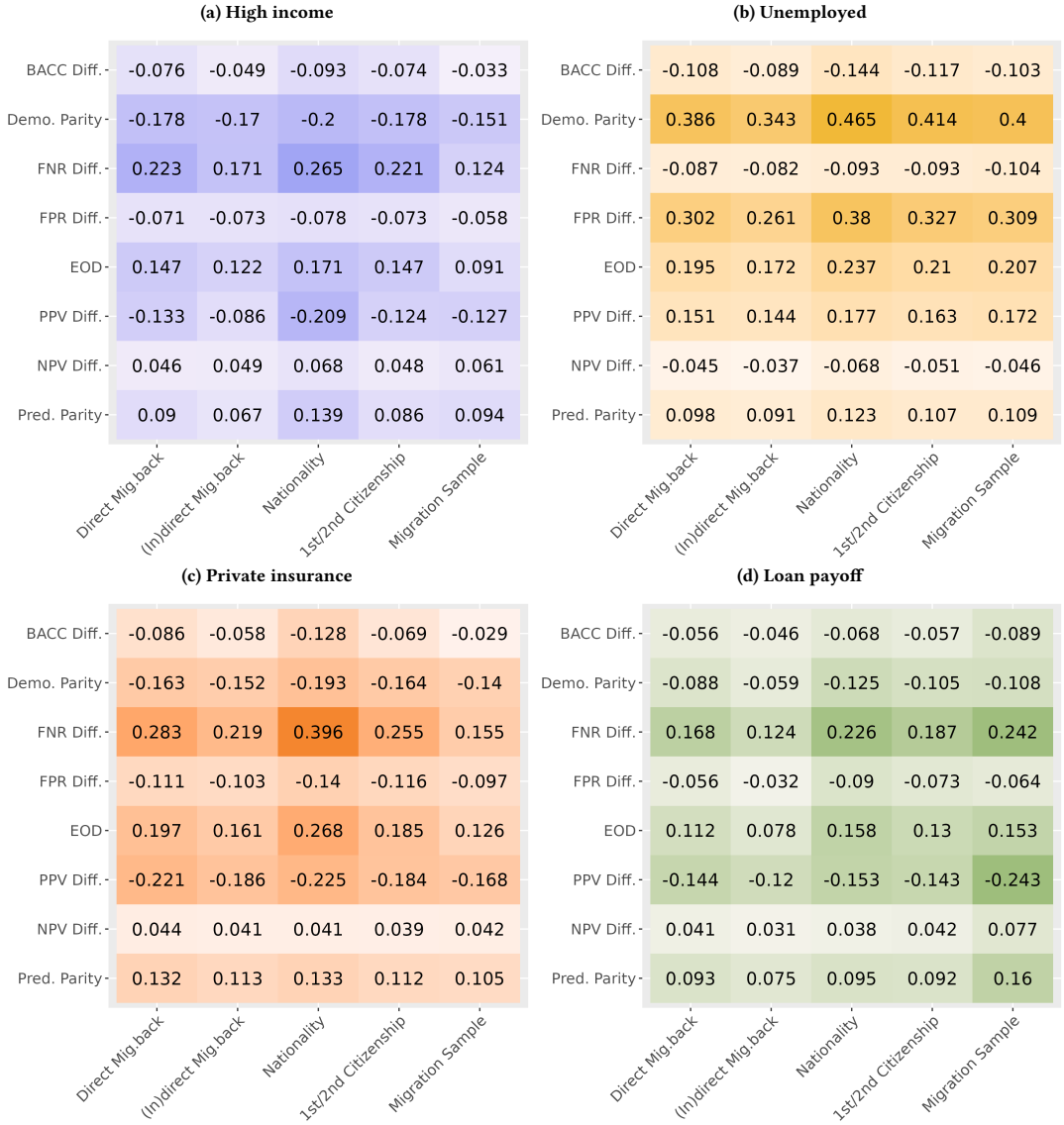

Figure 2: Fairness scores for the same prediction model can vary greatly between different ethnic classifications. Plots present fairness metrics computed for prediction models of four outcomes variables when defining the protected group based on different ethnic classifications. Each subplot shows scores for the same model and differences across columns can solely be attributed to differences in measuring ethnicity. Average scores over 10 model runs are reported.

While functional measures of ethnicity may be acceptable in arenas dominated by institutional discrimination, biases due to social processes may only be captured by more inclusive measures that recognize the complexities and multidimensionality of ethnicity.

Our empirical case study shows that the choice of ethnic classifications affects fairness evaluations in ADM. Using German survey data with diverse sets of minority populations, we demonstrate how prediction models across ADM domains (labor market, health, finance) (1) underperform for ethnic minorities in comparison to the majority population, and (2) display different ethnic disparities dependent on how ethnicity is operationalized and measured. The most pronounced unfairness scores were recorded when the protected group was defined based on the restrictive measure "nationality". However, strong disparities were also observed when defining ethnicity based on "direct migration background" or belonging to a migration subsample. If rigid thresholds are employed in fairness audits (e.g., the 80/20 rule for disparate impact), such differences can lead the same model to be either accepted or rejected given the specific operationalization of ethnicity that is implemented.

Our results underscore the importance of considering which ethnic classification is appropriate in a given context. While more

strict measures such as "nationality" may capture the starkest differences, it leaves out individuals who experience (algorithmic) disadvantages given their migration history or their family's biographies and experiences. In fact, as all five ethnic classifications in our fairness evaluations led to non-trivial unfairness scores, we propose that ADM applications should employ both restrictive and inclusive measures of ethnicity in fairness evaluations, next to careful considerations of how ethnic groups may have been subject to different manifestations of discrimination in the given application context and how each form of discrimination links to the measures of ethnicity that are constructed.

We further urge the research community to report and clarify the ethnic classifications employed in their studies. While we show variation based on different concepts of ethnicity, a single concept can also be measured in different ways, especially when it comes to complex dimensions such as culture or ethnic identity [27]. These challenges mirror the intricacies of operationalizing race [24] and extend to the delicate task of algorithmically inferring ethnic categories from other individual attributes [4, 37], which involves an implicit operationalization of ethnicity. We thus underscore the need for careful reflection and transparent reporting of the specific operationalization and measure(s) used.

Limitations. The ethnicity schemes presented in this paper do not constitute an exhaustive list of all possible approaches to ethnic categorization and analysis. Our discussion is intended to provide an overview of central concepts and dimensions that are employed in the study of ethnicity which may not fully capture the nuanced experiences of ethnic classifications in particular scenarios. Our empirical focus on Germany poses limitations, as it may not fully reflect the full spectrum of ethnic dynamics in Europe. The absence of an intersectional fairness analysis constrains the understanding of how intertwined social categorizations like ethnicity, gender and class jointly impact algorithmic discrimination. Despite these limitations, our study pioneers the evaluation of ADM fairness concerning ethnicity in a European setting. Hence, this study lays the groundwork for further research.

Building on the findings of this study, several avenues for future research can be explored, including the study of additional ethnic classifications across diverse contexts, considering ethnicity in ML tasks with non-tabular data, and intersectional fairness. By studying and measuring ethnicity in a principled way, researchers and policymakers can better understand distinct patterns of inequality, discrimination, and social exclusion that certain ethnic groups may face. It helps shed light on disparities in education, employment, healthcare, housing, and other areas that are increasingly becoming ADM domains, and informs efforts to address these inequalities.

6 RESEARCH ETHICS AND SOCIAL IMPACT

6.1 Ethical Considerations

The ethical concerns we encountered and addressed while conducting this research are:

(1) Ensuring the respectful and accurate representation of ethnic groups, avoiding the perpetuation of biases.
(2) Presenting a concise collection of dimensions and measures of ethnicity that reflect both the concept's inherent complexity and the (simplifying) implementations used in data collection practice.
(3) Handling data ethically, especially when it contains sensitive information like ethnic classifications. We made a formal request to obtain the SOEP data, a process that involved adherence to data protection standards. Upon receiving the data, we signed a disclosure agreement, committing to not share the data and to use it solely for the intended research purposes. In doing so, we respect the privacy and integrity of the individuals represented in the dataset.

6.2 Adverse Impact Statement

Potential adverse impacts of our research are:

(1) The misuse or misinterpretation of our findings, where the research might be taken out of context or used to justify the use of a specific measure of ethnicity that may be too restrictive for the given application domain.
(2) The potential over-generalization of our results. Our analysis is based on a specific case within the European context, and it might not be appropriate to view our findings as universally applicable across the entire European region.
(3) Our focus on ethnic classifications might reinforce the notion of rigid ethnic categories, which can be problematic as it ignores the fluidity and subjective nature of ethnic identity.

6.3 Researcher Positionality

Our team's individual backgrounds and affiliations informed this research. One author, born in Argentina with migrant grandparents, later relocated to Europe. Studying in Latin America before working in Germany has provided this author with a unique insight into the various ways ethnicity is perceived and operationalized in different cultural and social contexts. The authors were affiliated with an European organization during the work on this project which inherently motivated its topic and data used in this study. By publishing this work, we aim to foster a responsible dialogue on algorithmic fairness in Europe, particularly in relation to ethnic classifications. Our goal is to contribute to a more inclusive and ethically sensible approach in technology, recognizing the diverse and dynamic nature of society.

ACKNOWLEDGMENTS

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research, the Munich Center for Machine Learning and the Volkswagen Foundation, grant "Consequences of Artificial Intelligence for Urban Societies (CAIUS)".

REFERENCES

[1] Amina A. Abdu, Irene V. Pasquetto, and Abigail Z. Jacobs. 2023. An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1324–1333. https://doi.org/10.1145/3593013.3594083
[2] Richard Alba. 1992. *Ethnic Identity: The Transformation of White America.* Yale University Press.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[4] Lisa P. Argyle and Michael Barber. 2024. Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records. *American Political Science Review* 118, 2 (2024), 1058–1066. https://doi.org/10.1017/S0003055423000229

[5] Peter J. Aspinall. 2002. Collective terminology to describe the minority ethnic population: The persistence of confusion and ambiguity in usage. *Sociology* 36, 4 (2002), 803–816.

[6] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 116–128. https://doi.org/10.1145/3442188.3445875

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. https://fairmlbook.org/

[8] Teanna Barrett, Quanze Chen, and Amy Zhang. 2023. Skin Deep: Investigating Subjectivity in Skin Tone Annotations for Computer Vision Benchmark Datasets. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1757–1771. https://doi.org/10.1145/3593013.3594114

[9] Sebastian Benthall and Bruce D. Haynes. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 289–298. https://doi.org/10.1145/3287560.3287575

[10] Richard Berthoud. 1998. Defining ethnic groups: Origin or identity? *Patterns of Prejudice* 32, 2 (1998), 53–63. https://doi.org/10.1080/0031322X.1998.9970255

[11] Leo Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[12] Robin N. Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning Equitable Speech Technologies for Black Older Adults. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 379–388. https://doi.org/10.1145/3593013.3594005

[13] Jonathan Burton, Alita Nandi, and Lucinda Platt. 2010. Measuring ethnicity: challenges and opportunities for survey research. *Ethnic and Racial Studies* 33, 8 (2010), 1332–1349. https://doi.org/10.1080/01419870903527801

[14] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287594

[15] Lingwei Cheng, Isabel O Gallegos, Derek Ouyang, Jacob Goldin, and Dan Ho. 2023. How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 667–686. https://doi.org/10.1145/3593013.3594034

[16] Stephen Cornell and Douglas Hartmann. 2006. *Ethnicity and race: Making identities in a changing world*. Sage Publications.

[17] G.R. de Groot and O. Vonk. 2018. Acquisition of Nationality by Birth on a Particular Territory or Establishment of Parentage: Global Trends Regarding Ius Sanguinis and Ius Soli. *Netherlands International Law Review* 65 (2018), 319–335. https://doi.org/10.1007/s40802-018-0118-5

[18] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2022. Retiring Adult: New Datasets for Fair Machine Learning. arXiv:2108.04884 [cs.LG]

[19] European Social Survey. 2020. ESS 1-9, European Social Survey Cumulative File, Study Description. Bergen: NSD - Norwegian Centre for Research Data for ESS ERIC.

[20] Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in NLP tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1479–1492. https://doi.org/10.1145/3593013.3594094

[21] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[22] Jan Goebel, Markus M. Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder, and Jürgen Schupp. 2019. The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik* 239, 2 (2019), 345–360. https://doi.org/10.1515/jbnst-2018-0022

[23] Stuart Hall and Paul du Gay. 1996. *Questions of Cultural Identity*. SAGE Publications.

[24] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[25] Anthony F. Heath, Catherine Rothon, and Elina Kilpi. 2008. The Second Generation in Western Europe: Education, Unemployment, and Occupational Attainment. *Annual Review of Sociology* 34, 1 (2008), 211–235. https://doi.org/10.1146/annurev.soc.34.040507.134728

[26] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Gender and Racial Bias in Visual Question Answering Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1280–1292. https://doi.org/10.1145/3531146.3533184

[27] Jürgen H. P. Hoffmeyer-Zlotnik and Uwe Warner. 2010. *Measuring ethnicity in cross-national comparative survey research*. GESIS - Leibniz-Institut für Sozialwissenschaften, Bonn. https://doi.org/10.21241/ssoar.26123

[28] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. Robots Enact Malignant Stereotypes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 743–756. https://doi.org/10.1145/3531146.3533138

[29] Yasmin Hussain and Paul Bagguley. 2005. Citizenship, ethnicity and identity: British Pakistanis after the 2001 'riots'. *Sociology* 39, 3 (2005), 407–425. https://doi.org/10.1177/0038038505052493

[30] Kai Ingwersen and Stephan L Thomsen. 2021. The immigrant-native wage gap in Germany revisited. *The Journal of Economic Inequality* 19, 4 (2021), 825–854. https://doi.org/10.1007/s10888-021-09493-8

[31] Stefania Ionescu, Anikó Hannák, and Kenneth Joseph. 2021. An Agent-based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 412–423. https://doi.org/10.1145/3442188.3445904

[32] Gauri Kambhatla, Ian Stewart, and Rada Mihalcea. 2022. Surfacing Racial Stereotypes through Identity Portrayal. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1604–1615. https://doi.org/10.1145/3531146.3533217

[33] Selin Kara, Stefan Zimmermann, and SOEP Group. 2022. SOEPcompanion (v37). https://www.diw.de/documents/publikationen/73/diw_01.c.855721.de/diw_ssp1192.pdf

[34] Zaid Khan and Yun Fu. 2021. One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 587–597. https://doi.org/10.1145/3442188.3445920

[35] Rebecca Y. Kim. 2011. Religion and Ethnicity: Theoretical Connections. *Religions* 2, 3 (2011), 312–329. https://doi.org/10.3390/rel2030312

[36] I Kogan. 2006. Labor markets and economic incorporation among recent immigrants in Europe. *Social Forces* 85, 2 (2006), 679–721. https://www.jstor.org/stable/4494936

[37] Jeffrey W Lockhart, Molly M King, and Christin Munsch. 2023. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour* 7, 7 (2023), 1084–1095.

[38] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the Applicability of Machine Learning Fairness Notions. *SIGKDD Explorations Newsletter* 23, 1 (may 2021), 14–23. https://doi.org/10.1145/3468507.3468511

[39] Iain Mathieson and Aylwyn Scally. 2020. What is ancestry? *PLoS Genetics* 16, 3 (2020), 1–6. https://doi.org/10.1371/journal.pgen.1008624

[40] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2021). https://doi.org/10.1145/3457607

[41] David Miller. 1995. *On Nationality*. Oxford University Press, New York.

[42] Minnesota Population Center. 2020. Integrated Public Use Microdata Series, International: Version 7.3. https://doi.org/10.18128/D020.V7.3 Accessed: [December 1, 2023].

[43] Ann Morning. 2008. Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round. *Population Research and Policy Review* 27, 2 (4 2008), 239–272. https://doi.org/10.1007/s11113-007-9062-5

[44] Ann Morning. 2015. *Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round*. Springer International Publishing. https://doi.org/10.1007/978-3-319-20095-8

[45] Rainer Münz and Ralf E. Ulrich. 1998. Changing Patterns of Immigration to Germany, 1945-1997. https://migration.ucdavis.edu/rs/more.php?id=69

[46] European Commission: European Migration Network. 2022. EMN Asylum and Migration Glossary. https://home-affairs.ec.europa.eu/networks/european-migration-network-emn/emn-asylum-and-migration-glossary_en Accessed: [December 1, 2023].

[47] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 89. https://doi.org/10.1145/

3287560.3287593
[48] Council of European Union. 2000. Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation.
[49] Council of European Union. 2000. Racial Equality Directive. Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.
[50] Council of European Union. 2004. Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services.
[51] Council of European Union. 2007. Charter of Fundamental Rights of the European Union.
[52] Council of European Union. 2012. Treaty on European Union.
[53] Council of European Union. 2012. Treaty on the Functioning of the European Union.
[54] Office of Management and Budget. 2024. Revisions to OMB's Statistical Policy Directive No. 15: Standards for maintaining, collecting, and presenting federal data on race and ethnicity. https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and. Accessed April 26, 2024.
[55] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–16. https://doi.org/10.1145/3313831.3376392
[56] Michael Omi and Howard Winant. 2014. *Racial Formation in the United States* (3 ed.). Routledge.
[57] Carsten Orwat. 2019. Diskriminierungsrisiken durch Verwendung von Algorithmen: Eine Studie, erstellt mit einer Zuwendung der Antidiskriminierungsstelle des Bundes. *Nomos* (2019).
[58] Oxford English Dictionary. 2023. *s.v. "ethnicity (n.), sense 2"*. https://doi.org/10.1093/OED/5152233166 Accessed December 15, 2023.
[59] Oxford Learner's Dictionaries. 2023. s. v. "culture" (n.). https://www.oxfordlearnersdictionaries.com/definition/english/culture_1. Accessed December 2, 2023.
[60] Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. 2022. Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 973–987. https://doi.org/10.1145/3531146.3533159
[61] Autoriteit Persoonsgegevens. 2020. Belastingdienst/Toeslagen: De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag. , 84 pages. https://autoriteitpersoonsgegevens.nl/uploads/imported/onderzoek_belastingdienst_kinderopvangtoeslag.pdf
[62] Jean S Phinney, Irma Romero, Monica Nava, and Dan Huang. 2001. The role of language, parents, and peers in ethnic identity among adolescents in immigrant families. *Journal of youth and Adolescence* 30 (2001), 135–153. https://link.springer.com/article/10.1023/A:1010389607319
[63] Fred L Pincus. 1996. Discrimination comes in many forms: Individual, institutional, and structural. *American Behavioral Scientist* 40, 2 (1996), 186–194. https://doi.org/10.1177/0002764296040002009
[64] R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[65] Annette C. Scherpenzeel and Marcel Das. 2011. "True" longitudinal and probability-based Internet panels: Evidence from the Netherlands. In *Social and behavioral research and the Internet: Advances in applied methods and research strategies*, Manfred Das, Peter Ester, and Lars Kaczmirek (Eds.). Routledge/Taylor & Francis Group, 77–104.
[66] Patrick Simon. 2012. Collecting ethnic statistics in Europe: a review. *Ethnic and Racial Studies* 35, 8 (2012), 1366–1391. https://doi.org/10.1080/01419870.2011.607507
[67] Patrick Simon. 2015. The Choice of Ignorance: The Debate on Ethnic and Racial Statistics in France. In *Social Statistics and Ethnic Diversity*, Patrick Simon, Victor Piché, and Antoine Gagnon (Eds.). Springer, Cham, Chapter 4. https://doi.org/10.1007/978-3-319-20095-8_4
[68] Patrick Simon. 2017. The failure of the importation of ethno-racial statistics in Europe: debates and controversies. *Ethnic and Racial Studies* 40, 13 (2017), 2326–2332. https://doi.org/10.1080/01419870.2017.1344278
[69] Ana Valdivia and Martina Tazzioli. 2023. Datafication Genealogies beyond Algorithmic Fairness: Making Up Racialised Subjects. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 840–850. https://doi.org/10.1145/3593013.3594047
[70] Marvin van Bekkum and Frederik Zuiderveen Borgesius. 2021. Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security* 23, 4 (2021), 323–340. https://doi.org/10.1177/13882627211031257
[71] Gert G Wagner, Joachim R Frick, and Jürgen Schupp. 2007. The German socio-economic panel study (SOEP)-Scope, evolution and enhancements. *Journal of Contextual Economics–Schmollers Jahrbuch* 1 (2007), 139–169.
[72] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. 2023. Algorithmic Unfairness through the Lens of EU Non-Discrimination Law Or Why the Law is not a Decision Tree. *ACM International Conference Proceeding Series* (2023), 805–816. https://doi.org/10.1145/3593013.3594044
[73] Charles Westin. 2003. Young People of Migrant Origin in Sweden. *The International Migration Review* 37, 4 (2003), 987–1010. http://www.jstor.org/stable/30037783
[74] Anne Kathrin Will. 2019. The German statistical category "migration background": Historical roots, revisions and shortcomings. *Ethnicities* 19, 3 (2019), 535–557. https://doi.org/10.1177/1468796819833437
[75] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. 2022. Evidence for Hypodescent in Visual Semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1293–1304. https://doi.org/10.1145/3531146.3533185
[76] Robert Wolfe and Aylin Caliskan. 2022. Markedness in Visual Semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1269–1279. https://doi.org/10.1145/3531146.3533183

A DATA AND VARIABLES

The German Socio-Economic Panel study (SOEP) is one of the most important data sources for empirical research in sociology, economics, psychology and political science in Europe [22, 71]. Since its introduction in 1984, annual surveys of the German population on a wide range of topics have been conducted following rigorous survey methodology. Its initial sample has been continuously complemented and refreshed to account for population dynamics, ensure sufficient coverage of specific subpopulations, and to enlarge sample sizes. Given its broad topical scope and detailed coverage of minority populations, the SOEP has great potential for fair ML research. We exemplify its usage for fairness evaluations in the code that accompanies this paper, which can serve as a starting point for future research. Individuals can apply for data access with the German Institute for Economic Research (DIW).

The following tables provide more detail on the data we used in our case study, including our selection of predictor variables (Table 3) and descriptive statistics for the outcomes (Table 4a), ethnic classifications (Table 4b) and predictors used in all four prediction tasks (Table 5).

B ADDITIONAL RESULTS

Table 3: Predictors used for each prediction task

Predictor	High Income	Unemployed	Private Insurance	Loan Payoff
Age	✓	✓	✓	✓
Sex	✓	✓	✓	✓
Years of Education	✓	✓	✓	✓
Marital Status	✓	✓	✓	✓
Type of Household	✓	✓	✓	✓
State of Residence	✓	✓	✓	✓
Employment Status	✓		✓	✓
Disability			✓	
Working Experience (Full-Time)	✓			✓
Working Experience (Part-Time)	✓			✓
Unemployment Experience		✓		✓

Table 4: Summary statistics

(a) Outcome variables

Variable	N	Percent
High Income	17438	
... high	4259	24%
... low	13179	76%
Unemployed	28910	
... no	25933	90%
... yes	2977	10%
Private Insur.	25712	
... no	22722	88%
... yes	2990	12%
Loan Payoff	29050	
... no	22648	78%
... yes	6402	22%

(b) Ethnic classifications

Variable	N	Percent
Direct Mig.back	29888	
... no	21734	73%
... yes	8154	27%
(In)direct Mig.back	29888	
... no	19903	67%
... yes	9985	33%
Nationality	29888	
... no	23487	79%
... yes	6401	21%
1st/2nd Citizenship	29888	
... no	22465	75%
... yes	7423	25%
Migration Sample	29888	
... no	22629	76%
... yes	7259	24%

Table 5: Summary statistics of predictor variables

Variable	N	Mean/ Percent	Std. Dev.	Min	Max
Age	29888	46.6	18	17	103
Years of Education	27789	12.0	2.9	7	18
Working Experience (Full-Time)	29587	15.1	14	0	58
Working Experience (Part-Time)	29587	3.7	6.6	0	52
Unemployment Experience	29587	1.3	3.2	0	49
Sex	29888				
... [1] Male	14232	48%			
... [2] Female	15656	52%			
Marital Status	29767				
... [1] Married, living together	16782	56%			
... [2] Married, permanently separated	759	3%			
... [3] Single	8051	27%			
... [4] Divorced / registered partnership annulled	2569	9%			
... [5] Widowed / life partner from registered partnership deceased	1396	5%			
... [6] husband/wife abroad	132	0.4%			
... [7] Registered partnership, living together	65	0.2%			
... [8] Registered partnership, living separately	13	0.04%			
Type of Household	29888				
... [1] 1-Pers.-HH	4483	15%			
... [2] Couple Without Children	8156	27%			
... [3] Single Parent	2654	9%			
... [4] Couple With Children LE 16	7128	24%			
... [5] Couple With Children GT 16	3258	11%			
... [6] Couple With Children LE And GT 16	3075	10%			
... [7] Multiple Generation-HH	403	1%			
... [8] Other Combination	731	2%			
Employment Status	29879				
... [1] Full-Time Employment	10293	34%			
... [2] Regular Part-Time Employment	4375	15%			
... [3] Vocational Training	898	3%			
... [4] Marginal, Irregular Part-Time Employment	1772	6%			
... [5] Not Employed	12497	42%			
... [6] Sheltered workshop	44	0.1%			
Disability	25688				
... [1] Yes	3118	12%			
... [2] No	22570	88%			

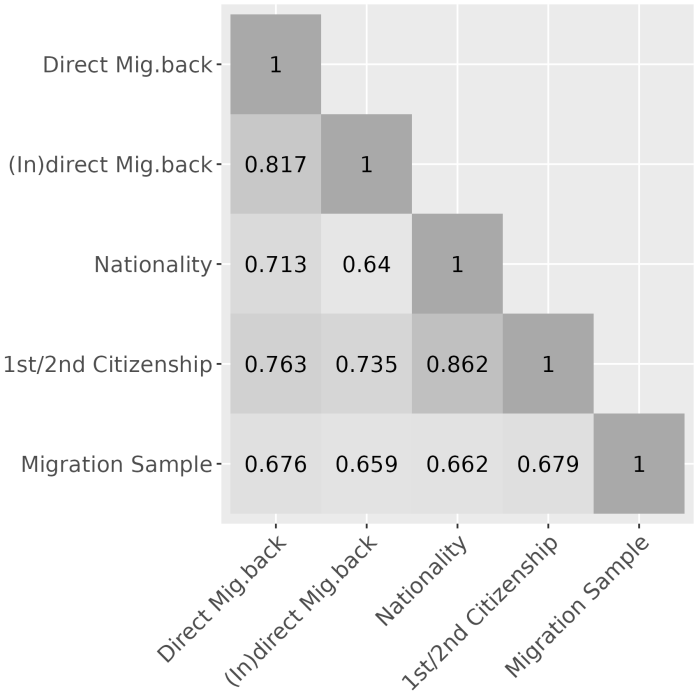

Figure 3: Different concepts of ethnicity capture different ethnic subpopulations. We present Jaccard similarities between ethnic classifications by calculating $J(S_a, S_b) = \frac{|S_a \cap S_b|}{|S_a|+|S_b|-|S_a \cap S_b|}$ with $|S_a|$, $|S_b|$ being the total number of individuals belonging to ethnic classifications a, b and $|S_a \cap S_b|$ being the number of individuals that are classified as being of ethnic origin according to both classifications.

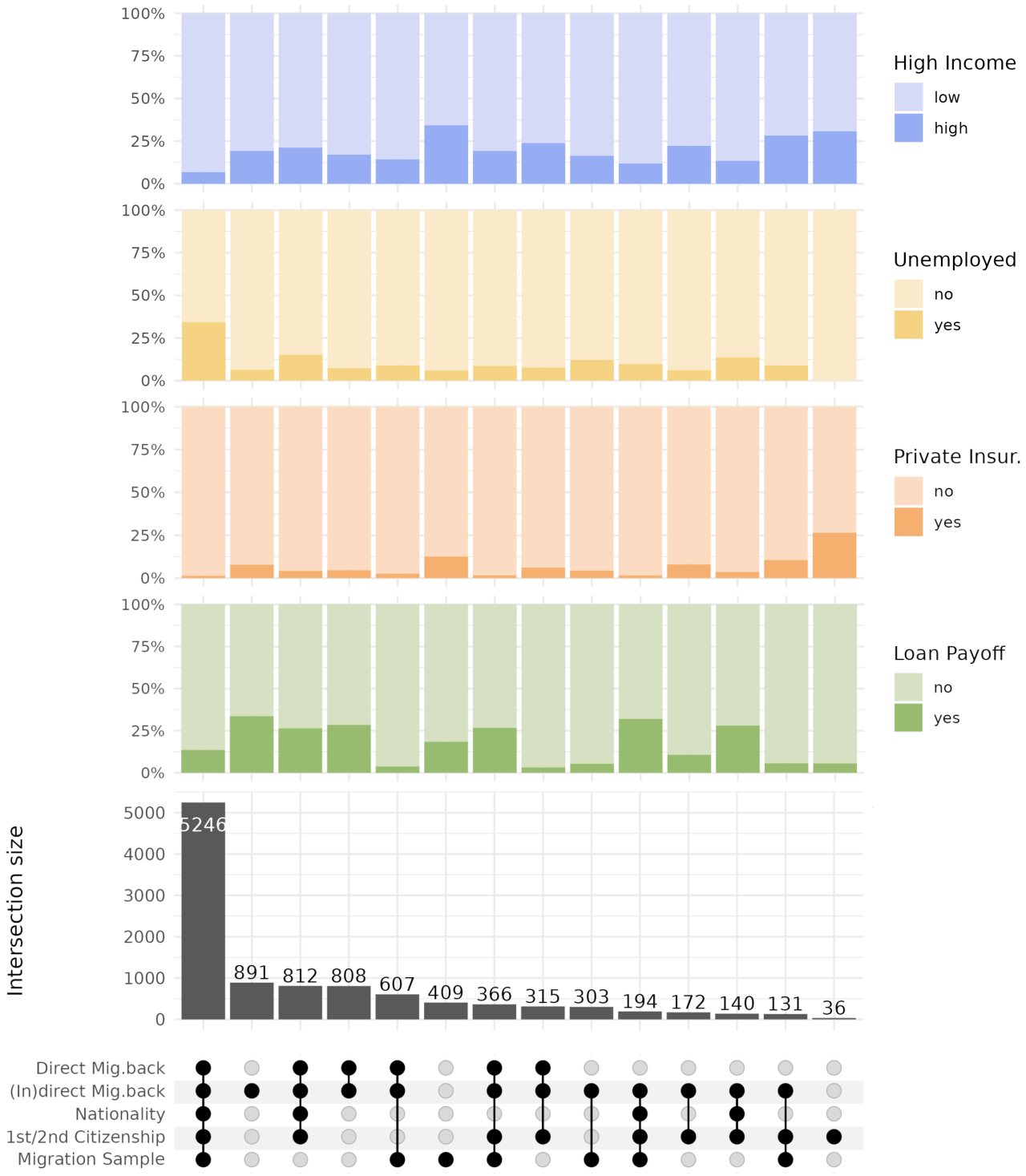

Figure 4: Sizable ethnic populations are only identifiable by distinct measures. Lower part: Intersection plot showing distinct (exclusive) subpopulations based on combinations of ethnicity measures and their set size. Upper part: Base rate plots for each outcome variable and set intersection.

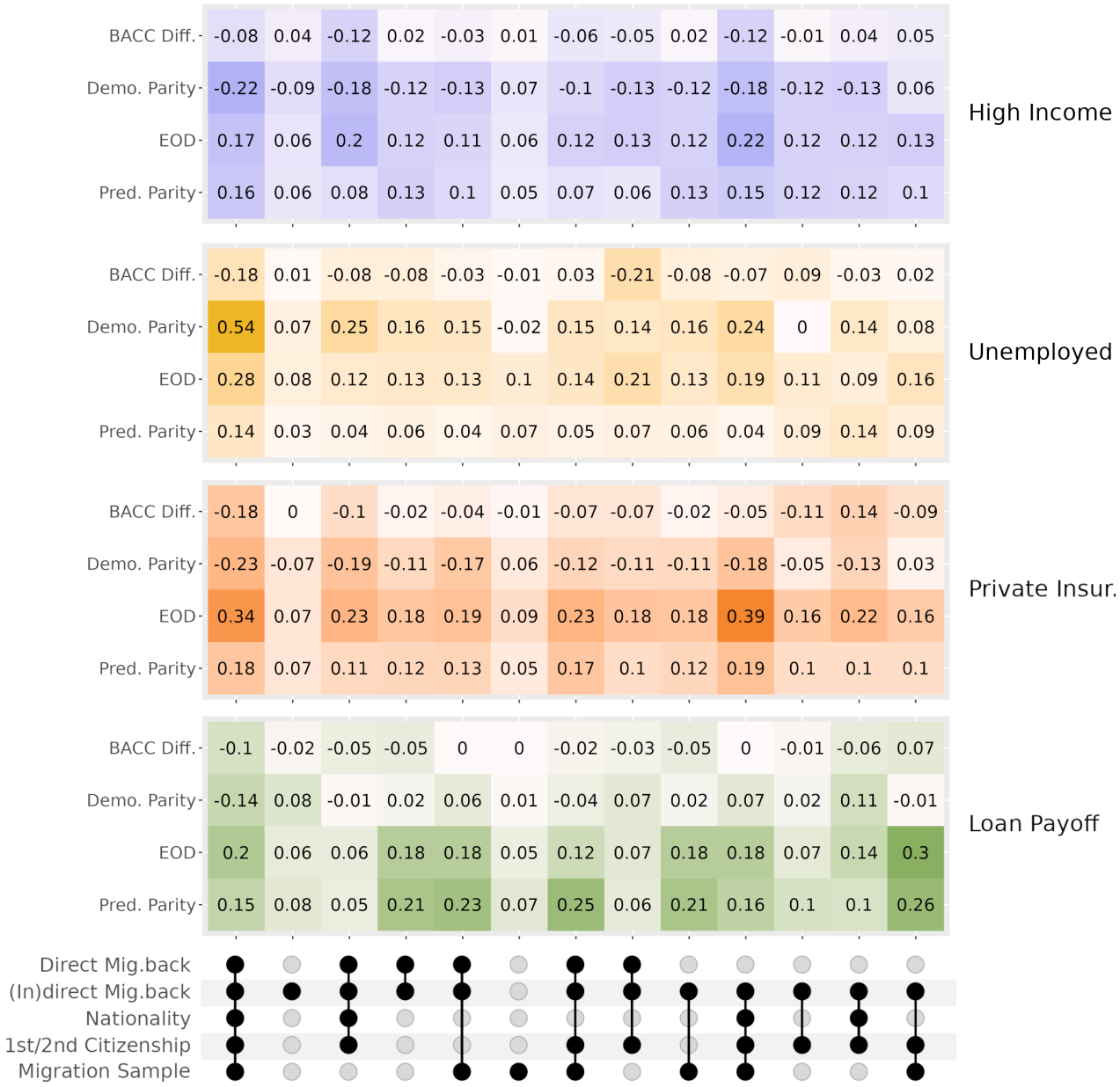

Figure 5: Intersections of ethnic classifications identify distinct unfairness patterns. Lower part: Intersection plot showing distinct (exclusive) subpopulations based on combinations of ethnicity measures. Upper part: Fairness scores for each outcome variable and set intersection. The reference (unprotected) group in all comparisons are individuals who are not classified as migrants by any measure of ethnicity. Average scores over 10 model runs are reported.