4-17-2023

Name-based demographic inference and the unequal distribution of misrecognition

Jeffrey W. Lockhart

Molly M. King
Santa Clara University, mmking@scu.edu

Christin Munsch

Recommended Citation

Lockhart, J. W., King, M. M., & Munsch, C. (2023). Name-based demographic inference and the unequal distribution of misrecognition. Nature Human Behaviour, 7(7), 1084–1095. https://doi.org/10.1038/s41562-023-01587-9

Name-Based Demographic Inference and the Unequal Distribution of Misrecognition

Jeffrey W. Lockhart[*,1], Molly M. King[2], and Christin Munsch[3]

[*] jlockhart@uchicago.edu

[1] Department of Sociology, University of Chicago, Chicago, USA
[2] Department of Sociology, Santa Clara University, Santa Clara, USA
[3] Department of Sociology, University of Connecticut, Storrs, USA

6 April, 2023

This is an author manuscript of an article that has been accepted for publication in *Nature Human Behavior*. Suggested citation:

Lockhart, Jeffrey W., King, Molly M., and Munsch, Christin. 2023. "Name-Based Demographic Inference and the Unequal Distribution of Misrecognition." *Nature Human Behavior*. xx(xx):xx-xx. DOI: 10.1038/s41562-023-01587-9

Name-Based Demographic Inference and the Unequal Distribution of Misrecognition

Abstract

Academics and companies increasingly draw on large datasets to understand the social world, and name-based demographic ascription tools are widespread for imputing information like gender and race that are often missing from these large datasets. These approaches have drawn criticism on ethical, empirical, and theoretical grounds. Employing a survey of all authors listed on articles in sociology, economics, and communications journals in the Web of Science between 2015 and 2020, we compared self-identified demographics with name-based imputations of gender and race/ethnicity for 19,924 scholars across four gender ascription tools and four race/ethnicity ascription tools. We find substantial inequalities in how these tools misgender and misrecognize the race/ethnicity of authors, distributing erroneous ascriptions unevenly among other demographic traits. Because of the empirical and ethical consequences of these errors, scholars need to be cautious with the use of demographic imputation. We recommend five principles for the responsible use of name-based demographic inference.

Keywords: Gender; Misgendering; Racial Misrecognition; Technology; Name-Based Inference

The digital age has made large data sets easily accessible, including databases with thousands of newspapers, millions of academic publications, or billions of social media posts. But these data generally lack demographic variables like gender, race/ethnicity, class, age, and religion that are at the core of traditional social research and marketing applications. They do, however, contain people's (real or screen) names. Concequently, name-based demographic inference is widespread in both computational social science and private industry. Practitioners take a name like "Adam" and impute "male" and a name like "Smith" and impute "British-Origin" or "non-Hispanic white." Popular tools for gender imputation such as genderize.io, M3-Inference, and R's gender and predictrace packages have a collective 945 citations in Google Scholar. Several have been commercialized for market research, app developers, and other private uses. Related tools like ethnicolor, predictrace, and WRU exist for inferring race/ethnicity from names, and still other tools have been created for age and religion. Academics-including one of these authors-have used these tools to shed light on gender and racial inequality in science, journalism, and online communities.[1–4] However, the tools have also drawn criticism from scholars both for ethical and validity concerns including offense to identity, the reification of gender binaries, and potentially conclusions.[5–8]

Efforts to evaluate the accuracy of name-based demographic inference typically involve relatively modest sample sizes, few covariates, and most importantly, human guessing as the ground truth.[9] They test, for example, whether machine guesses align with guesses from other humans. This approach fails to address the gap between gender identity and ascribed gender, and ignores the importance of covariates like nationality, race/ethnicity, and class which are known to affect naming.[10,11] We advance the literature on both fronts. First, we elaborate the gap between ascribed identities and other aspects of gender and race. Then, moving beyond the

question of overall accuracy we ask for whom these tools are more or less accurate, and thus who is systematically advantaged, harmed, or erased by these technologies. Rather than seeking to find a tool with the best performance or making claims about universal error rates, we argue that the fundamentally ambiguous linguistic and cultural processes of naming necessarily result in heterogeneous error rates. Analyses with different tools or populations will have different distributions of errors, but the fundamental ambiguity and heterogeneity we show in the relationship between naming and demographic labels is inescapable.

Drawing on a survey of 19,924 authors of social science journal articles, we examine gender and racial misclassification in a trans- and nonbinary-inclusive way along with nationality, sexuality, disability, parental education, and name-changes. By combining names from a database of publications without demographic data—the kind these tools are often used for—with original surveys of self reported demographic data, we can investigate errors in name-gender and name-race imputation.

Results show an overall error rate for gender prediction of 4.6% in our sample using the most popular tool, genderize.io. However, there are drastic differences in the error rate by subgroup. By definition, automated gender inference is wrong for all 139 nonbinary scholars in our sample. The algorithm was wrong 3.5 times more often for women than men, and some subgroups like Chinese women have error rates over 43%. For scientists, these disparities will bias results and inferences . For the subjects, misgendering and misclassification of race/ethnicity can produce significant harms, ethical implications of which are heightened by the unequal distribution of harm across groups.[5,6,12]

Disparities in error rates are fundamental problems with the information content of names and the cultural construction of gendered and racialized groups. Thus they cannot be eliminated with more data or better statistics.[13] They can, however, suggest substantively interesting insights about the world. For example, Black respondents whose parent(s) have a PhD are more likely to be labeled Black by the algorithm than those whose parents did not attend college, suggesting that highly educated Black people may be more likely to give their children distinctively Black names, or that first generation Black scholars may have a harder time succeeding with distinctively Black names than their colleagues with academic parents. Yet, the reverse is true among Indian scholars, suggesting that highly educated people from India may give their children less distinctively Indian names. Only by attending to variation among and within groups will scholars be able to understand the validity of their measures and the social processes of gendering and racialization.

Based on our findings, we recommend five principles for conducting name-based demographic inference. Which of these is most appropriate and practical depends on the nature of the data and the inquiry. First, in cases where name-based demographic inference may not be theoretically or ethically justified, we urge critical refusal. Second, when perceived gender or race/ethnicity is of interest, then measures of demographic inference are warranted. Third, inference can be shaped to be specific to the researcher's population of interest using domain expertise. Fourth, exert caution by deploying name-based imputation only for subgroups with high accuracy and consistency. And, fifth, name-based demographic estimates can be used better in aggregate measures than individual classifications.

In what follows, we first motivate our work by discussing why demographics are correlated with names. Next, we review methods and limitations of imputation, before focusing on misgendering and misrecognizing race/ethnicity and the consequences thereof.

Gender is socially constructed. Behaviors, sounds, and objects take on and change gendered associations as part of cultural meaning-making .[14] Similarly, people and things do not have racial essences; they are instead racialized by institutional, cultural, and interpersonal processes.[15] Nothing inherent in a sequence of characters or phonemes that makes up a name ties it to the gender, race, or class of the person it names. Nevertheless, people often name their children in ways that (consciously or unconsciously) signal gender, racial/ethnic, religious, and even class membership.[10,16–18] Other times, they resist these associations by choosing ambiguous names for their children[16,17], or by changing their own names later in life. The aggregate result of these choices is an imperfect cultural consensus around the gendered, racialized, and other associations of many names. What name-based demographic imputation tools measure, then, is not the "ground truth" of a person's or name's gender or race (which does not exist), but rather the cultural "consensus estimates of how each name is gendered" or racialized.[13]

Cultural consenses are necessarily local and contextual to specific populations. For example, in the contemporary US, the name "Andrea" typically refers to women, but in Italy, it typically refers to men. Other names, like "Leslie," are commonly used for both women and men, resulting in weaker demographic correlations and less social signaling information.[16,19]

Most name-based demographic imputation tools are simple naive-Bayes classifiers.[18,20] They start with a reference dataset of name-gender or name-race records like baby names from the US Social Security Administration and define the probability that a name belongs to each gender or racial group as the proportion of people with that name in each group in the reference data. If 77% of people named Leslie in the reference data are women, then each new Leslie is 77% likely to be a woman. Many turn this continuous probability into a discrete classification by selecting the gender or race with the highest probability. So, all Leslies would be labeled as women, and every man and nonbinary person with this name would be mislabeled, 100 - 77 = 23% of people (the Bayes error rate).

Some approaches use other features beyond whole names, like n-grams or geography,[18,21–26] potentially improving accuracy. Nevertheless information-theoretic limits mean the core problem remails (e.g., Leslies in Utah in 2015 have a different proportion of women than Leslies overall: some will still be misgendered).[13] Other approaches sacrifice overall accuracy in exchange for more equal error rates across groups by changing the classification thresholds.[20,26] Researchers interested in aggregate estimates rather than individual labels can improve performance by using the predicted probabilities (e.g., .77 woman) rather than discrete classifications (e.g., 1 = woman).

Critically, the reference data population is almost never the target population. This is trivially true: imputation is done because data lacks the variable. Reference data has the variable by definition. But it is true in a deeper sense as well: the populations these tools are typically used with (e.g. English language social science authors, people tweeting a specific hashtag, *Guardian* website commenters) are not common reference populations (e.g., people with social security numbers at birth; registered voters in Florida; the proprietary, black-box agglomeration of records scraped by genderize.io). These populations have different demographic distributions . The Sex and Gender section of the American Sociological Association, for instance, is 83% women, while the overall association is only 56% women.[27] If we use US Social Security Administration baby names-or even a sample of the ASA member database-as the reference data set, we are likely to underestimate the number of women in this section and overestimate it in sections like Mathematical Sociology (33% women).

Moreover, reference and target population often have different categories altogether. Nonbinary people write social science publications and tweets. But in terms of US federal administrative vital statistics, there are no nonbinary babies. Likewise, the administrative category "African American" cannot adequately represent the various categories by which people are racialized in Africa. There are also differences between populations in how people write names. Scientific publications are more likely to use initials; online trolls are more likely to use misleading pseudonyms or present fake identities; informal spaces are more likely to use shortened names or nicknames. All of these factors suggest higher and less predictable error rates for name-based demographic imputation.

These misrecognition errors can have important consequences. Humans automatically ascribe gender to one another, placing people into sex categories in of everyday interaction.[14,28] Without asking one's gender, . There is the possibility of *misgendering* – or ascribing a gender to someone that is incongruent with their sense of their own gender,which may or may not align with their chromosomes, genital configuration, or legal gender. Misgendering can cause a wide array of harm. Ascribing gender to people denies their agency and subjective experience of their own gender[5], especially when people deliberately name themselves to resist gender ascription (e.g., by selecting androgynous names or using initials), and deliberate misgendering has a long history as a tactic of bullying and harassment among cisgender people.[29,30] Misgendering is associated with adverse health outcomes[31] and experiencing violence.[32] This is especially common and harmful for trans people for whom misgendering carries added dimensions of existential weight and access to institutional resources like medical care and toilets.[33]

The automated systems we describe also ascribe gender to people, misgendering some fraction of them in the process. These systems operate on a larger scale, however, with different consequences. For example, ascribing demographic labels to people based on names raises ethical challenges central to the Belmont Report's principle, Respect for Persons.[34] (Indeed, people perceive misgendering as more harmful when it comes from algorithms than other humans.[35]) Moreover, some systems directly interact with the people they misgender–for example, automated systems and marketing materials that target persons for gendered products.[36] Others gatekeep physical space or institutional resources by automating access or influencing recommendations.[6] When people learn they have inadvertently misgendered someone, they tend to rely less on ascribed gender in the future. [37]We hope that the same will be true of people using name-based gender imputation.

Even when people are unaware that distant analysts are using automated systems to classify their gender, the ascriptions can be insidious. Such uses directly extend the long history of scientific and administrative actors exerting control over populations through gender classification, which is intimately bound up with colonial and eugenic projects.[6,36] The use of such systems may also reinforce beliefs that gender is binary, fixed, and knowable at a glance, which are empirically false,[6,38,39] and harmful to trans, nonbinary, intersex, and cisgender/endosex people[6.40] While such broader social harms are outside the purview of individual-focused research ethics frameworks, they remain important considerations for scientists.[41]

Like gender, race/ethnicity is a system of social categorization.[15] People racialize one another in everyday interaction and broader structural systems, and they have a stake in their own racial identities and how others perceive them. Of course, dominant racial categorization systems are more complex and category membership is more ambiguous than the dominant two-

category gender system. Some people are invested in having their race 'correctly' identified by others, some are deeply invested in passing in order to access legal, educational, and other freedoms they would otherwise be denied[42,43], or for the purpose of 'identity tourism.'[44] Nevertheless, racial categorization structures access to resources, exposure to violence, and other key dimensions of life. Moreover, colonial and eugenic projects of controlling populations by imposing categorizations upon them for scientific or administrative ends live on in automated race/ethnicity imputation systems. Additionally, outside perceptions influence one's sense of their own racial identity[45], further raising the stakes of racial classification technologies.

Of course, gender and race classification systems are not independent of one another, and neither are the technical systems designed to reproduce those classifications. Thus, attending to the intersections of identity in these systems aids in our understanding of the cultural and institutional processes that misattribute gender and race. For example, tools designed to classify gender from pictures of faces perform differently across groups, exhibiting the lowest accuracy with dark-skinned women.[46] Similarly, prior work without self-report data has shown that names from different parts of the world are misgendered at different rates, with European names misgendered least.[9] This produces ethical concerns as the benefits or harms of correct or incorrect classification are not evenly distributed. We explore further heterogeneity in error rates among algorithms designed for name-based demographic imputation.

Results

Our analyses reveal considerable heterogeneity in error rates for both gender and race imputation across demographic groups.

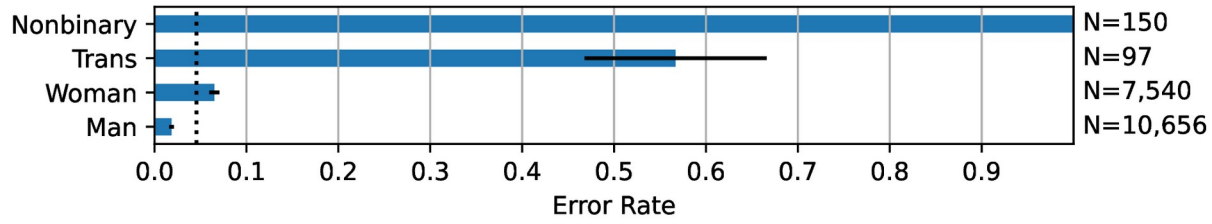

Figure 1: Proportion of people misgendered by gender when using genderize.io to label social science authors' gender. Error bars indicate 95% confidence intervals. The dotted line shows the population error rate (4.6%).

Misgendering

The relatively low overall error rates among the four algorithms tested-R's `gender` package had the lowest overall (4.4%) followed by genderize.io (4.6%)-obscure dramatic heterogeneity. We focus here on the most popular algorithm, genderize.io, but results for all algorithms show the same general pattern, (Figure S1). Error rates for men, women, trans, and nonbinary people are shown in Figure 1. Women are misgendered 3.5 times more often than men ($z=16.4$, $p=3.4\times10^{-60}$, $h=0.24$). Like other algorithms, by design genderize.io misgenders 100% of nonbinary people. The rate of misgendering for trans people is 57%.

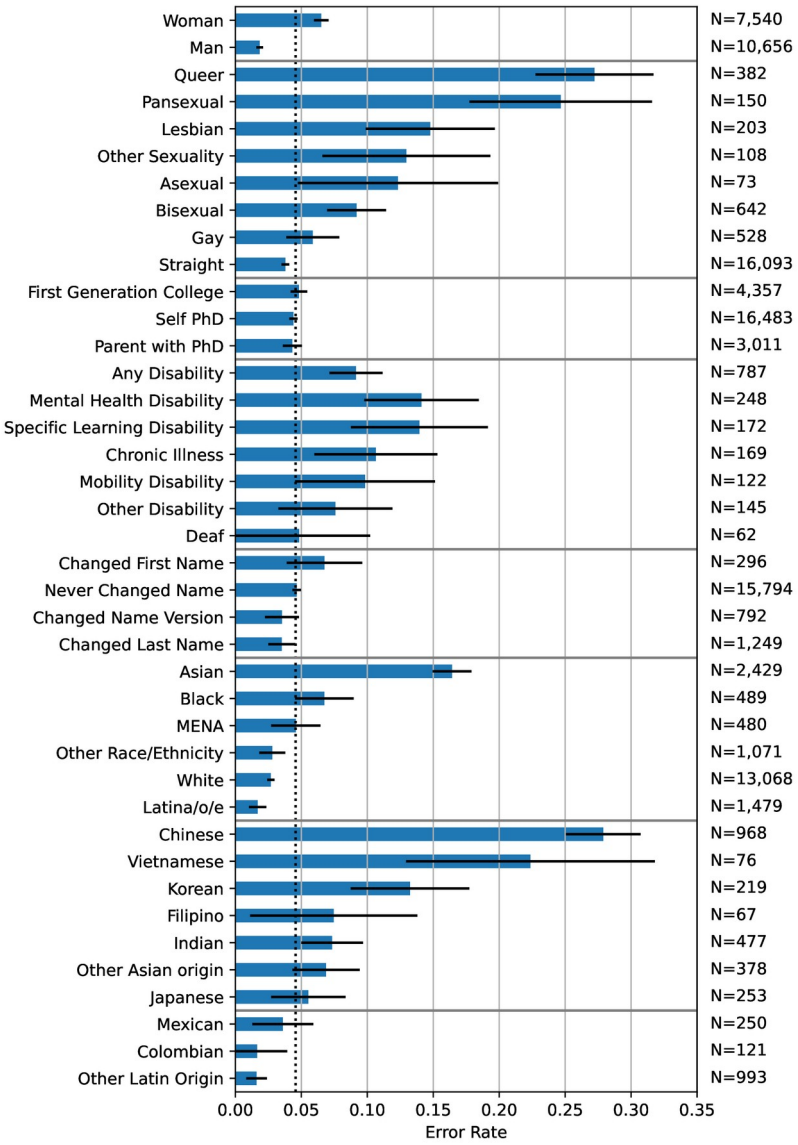

Figure 2: Proportion of people misgendered by sexuality, parental education, disability, name change history, race, and ethnicity when using genderize.io to label social science authors' gender. Error bars indicate 95% confidence intervals. The dotted line shows the population error rate (4.6%).

Misgendering is distributed unevenly along other demographic traits as well. Figure 2 shows rates of misgendering by sexuality, parental education, disability, name change history, and race/ethnicity. Notably, sexual minority people are misgendered more than their straight peers, as are people with disabilities, and Asian people. In contrast, white and Hispanic or Latina/o/e people are misgendered much less than other groups.

Yet not all sexual minorities are misgendered at the same rate: people with more marginal sexualities like queer and pansexual are misgendered much more often than gay and bisexual people. Likewise, within the broad US racial category "Asian," Chinese, Vietnamese, and to a

lesser extent Korean people, are misgendered much more often than Indian or Japanese people. We also observe variation among types of disability. We further decompose this inequality with a two-way cross-tabulation (Figure 3), revealing even more dramatic heterogeneity. For example, Chinese women are misgendered 43% of the time compared to Chinese men, who are misgendered 13% of the time ($z=11.1$, $p=6\times10^{-27}$, $h=0.32$). First name changes do not affect everyone equally: 1% of men who change their first name are misgendered, compared with 9% of women who do so ($z=2.7$, $p=0.0066$, $h=0.32$). That number is 69% for trans people of any gender (compared to cis people, $z=10.7$, $p=7.3\times10^{-27}$, $h=1.57$).

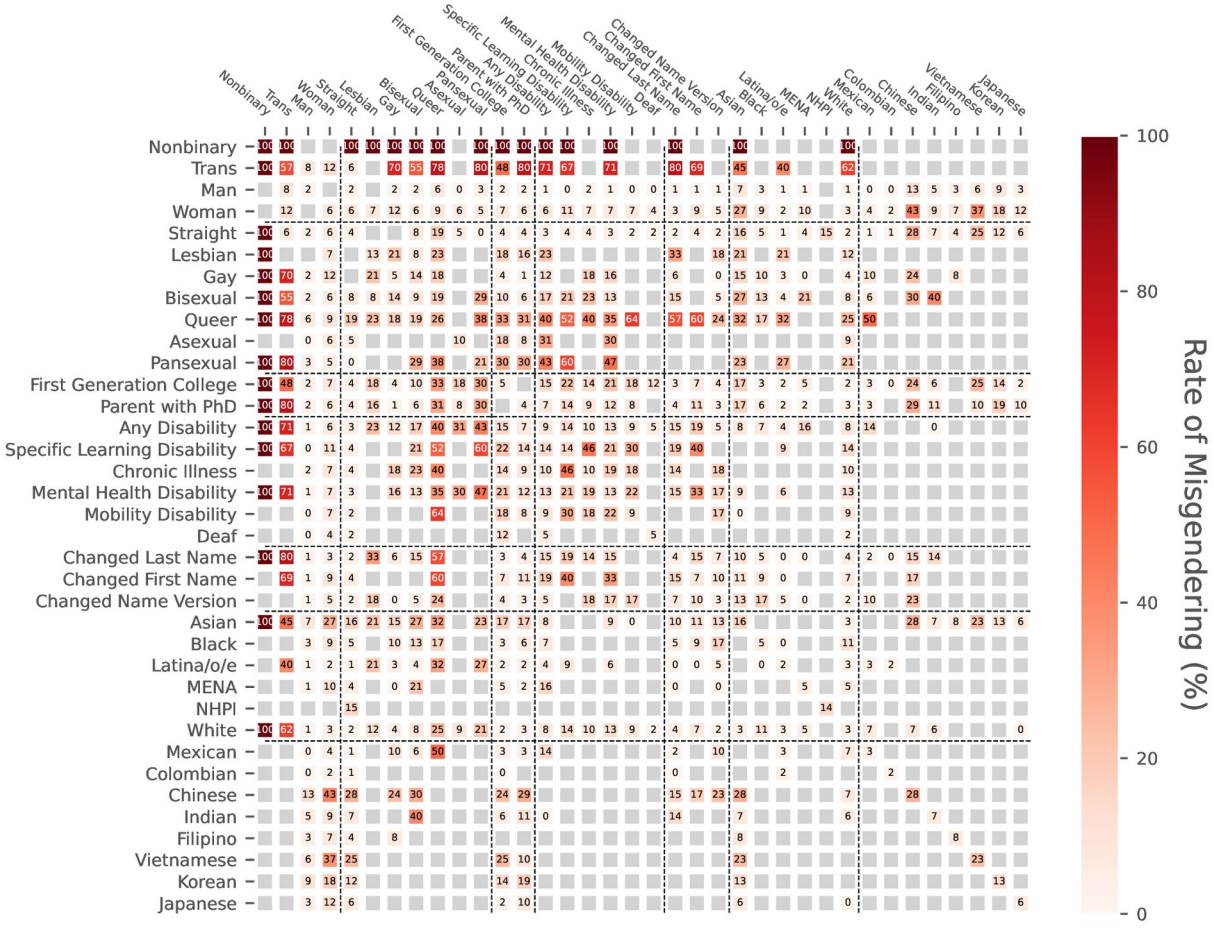

Figure 3: Two-way cross-tabulation of rates of misgendering by gender, sexuality, parental education, disability, name change history, race, and ethnicity when using genderize.io to label social science authors' gender. Numbers are percent misgendered. The top left corner shows that 57% of all trans people are misgendered and 8% of trans men are misgendered. Cells with fewer than 10 people are gray and not reported.

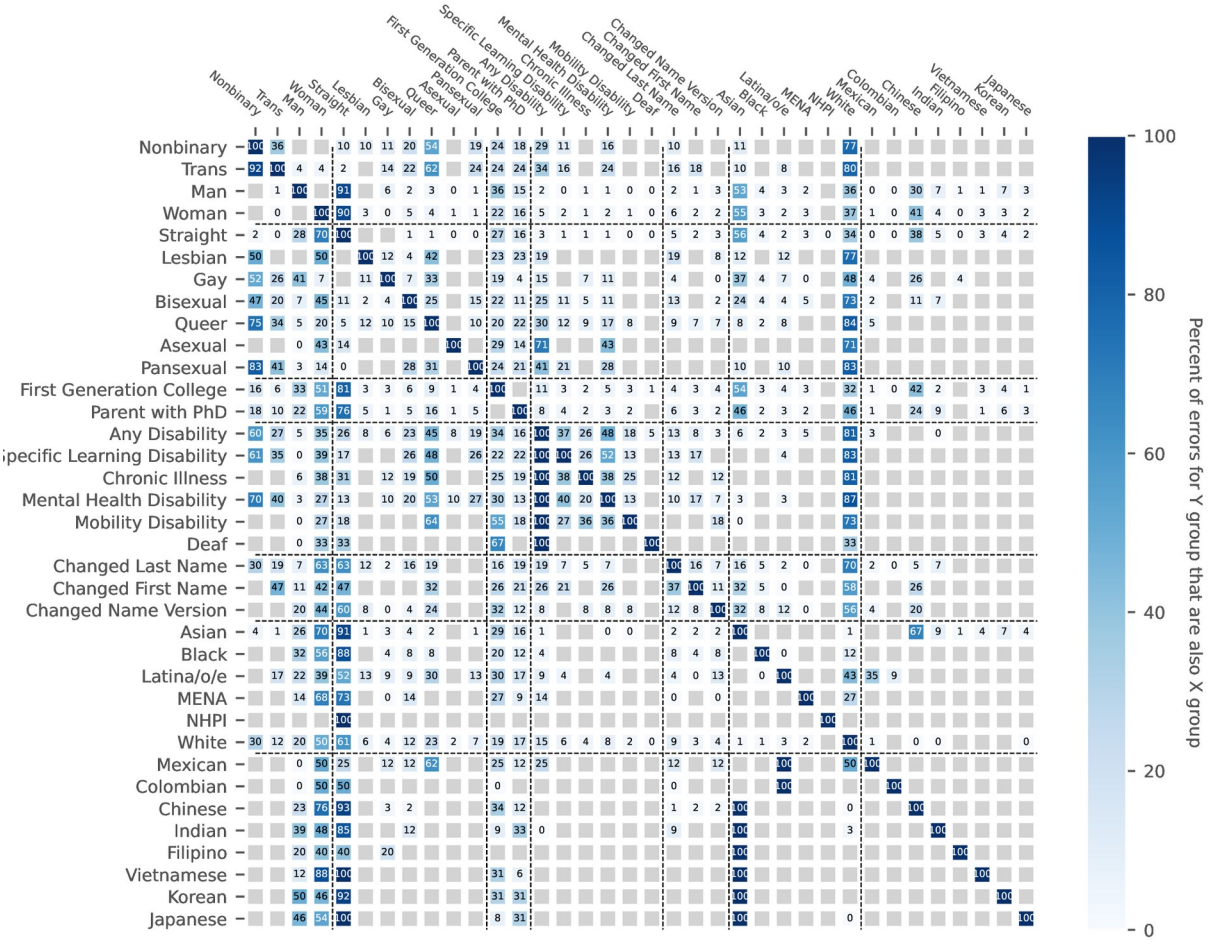

Figure 4: Apportionment of misgendering errors within groups using the genderize.io algorithm on social science authors. Numbers are percentages. The top left corner shows that 92% of trans people who are misgendered are also nonbinary, while only 36% of nonbinary people who are misgendered self-identify as trans.

These results are partly due to demographic confounding, underscoring our point: identities are not independently distributed in our population or any other. Figures S2 and S3 show the correlations and over-/under-representation among groups in our sample. Figure 4 shows the apportionment of errors within groups. The top left corner is instructive: 92% of trans people (the row) who are misgendered are also nonbinary (the column). Since nonbinary people are always misgendered, this means 92% of the errors for trans people are due to demographic overlap with nonbinary identity. Further down in the same column, we see that 52% of gay people, 60% of people with disabilities, and 30% of white people who are misgendered are nonbinary. In short, misgendering nonbinary people has spillover effects on accuracy in other demographic categories.

Spillover is not only a feature of nonbinary identity. For example, 88% of Vietnamese and 76% of Chinese people who are misgendered are women, in keeping with the overall higher rate of misgendering among women. This pattern, however, does not hold among Japanese

people, where 46% of those misgendered are women. This heterogeneity in both magnitude and direction of gender bias among subpopulations makes accounting for bias at the population level especially difficult.

Misrecognizing Race and Ethnicity

We conducted the same analyses for race/ethnicity. Notably, we use the most optimistic measure of accuracy in these analyses, counting even partially correct predictions as correct, to show that even by the most generous standards, the problem persists. Again, all algorithms have qualitatively similar results (Figure S4), and we focus on the best performing algorithm, R's `predictrace` package (Figure 5). Overall accuracies ranged from 47% to 86% when predicting broad US Census racial/ethnic categories of social science authors from their names. As expected, there is dramatic variation by race/ethnicity and national origin, with Black, Middle Eastern and North African, Filipino, and self-described "Other" misclassified between 55% and 80% of the time. By contrast, White, Asian, Chinese, Vietnamese, and Korean are mislabeled less than 10% of the time. Moreover, while we find little variation in the rate of racial misclassification by gender or disability, there is variation by both sexuality and name changes. Notably, sexuality is largely uncorrelated with race/ethnicity and national origin in our sample (see Figures S2 and S3), meaning demographic confounding is not the driving cause of sexual minorities' racial misclassification. Name changes, however, are weakly related. Changing one's name likely effects racial classification accuracy, for example when spouses adopt names with a different racial/ethnic association. Supporting this, 15% of racially misclassified women have published under different last names, compared to 7% of misclassified men ($z=6.3$, $p=2.2\times10^{-10}$, $h=0.28$). So while there is no significant overall difference between men's and women's racial misclassification rates ($z=0.55$, $p=0.58$, $h=0.009$), the factors driving these errors differ for each group.

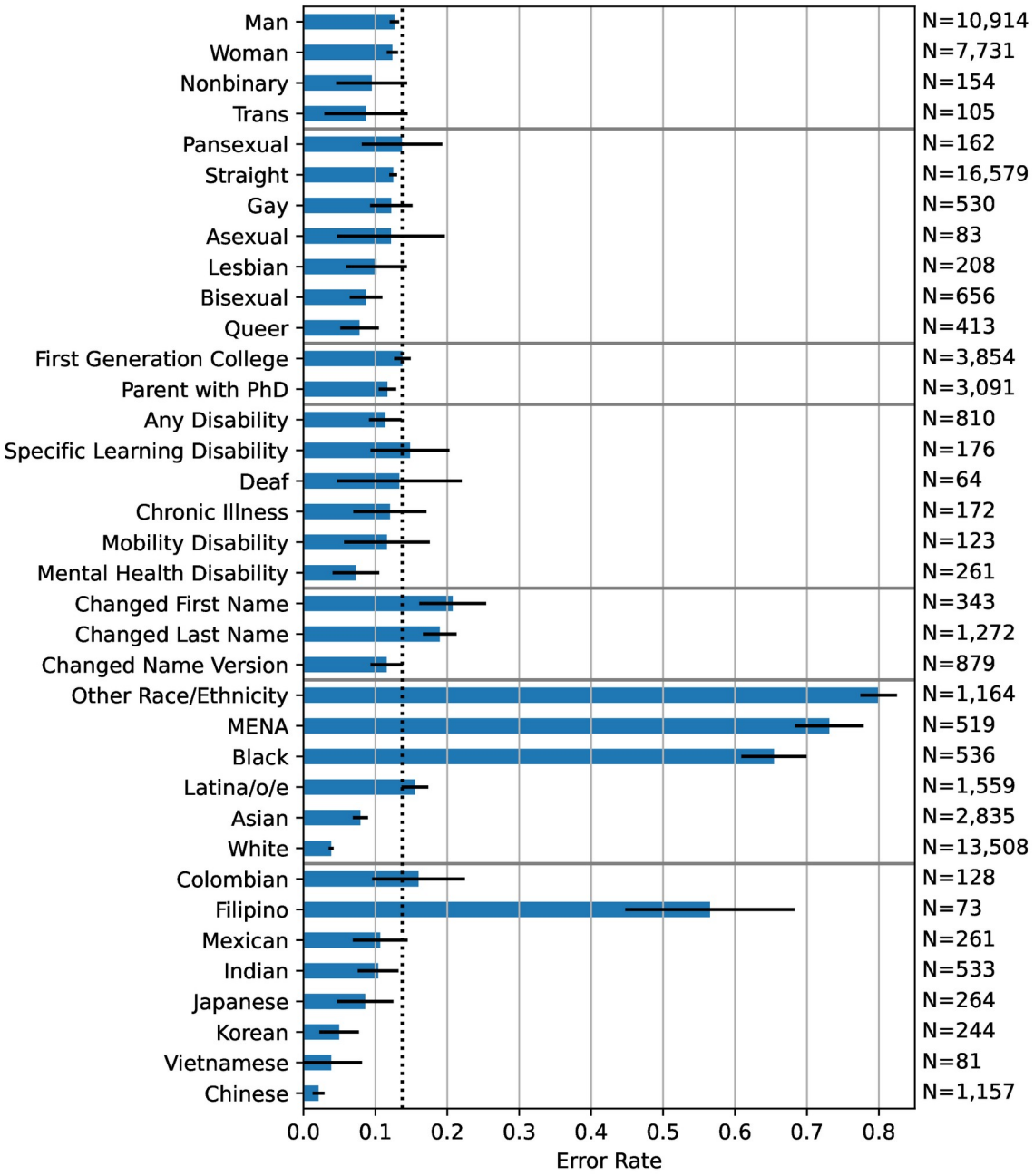

Figure 5: Proportion misclassified by race/ethnicity imputation using `predictrace` on social science authors. Error bars indicate 95% confidence intervals. The dotted line shows the overall error rate, 14%. Note the overall error rate is greater than the error rate for any gender because 628 people did not report a gender, and their race/ethnicity error rate is 51%.

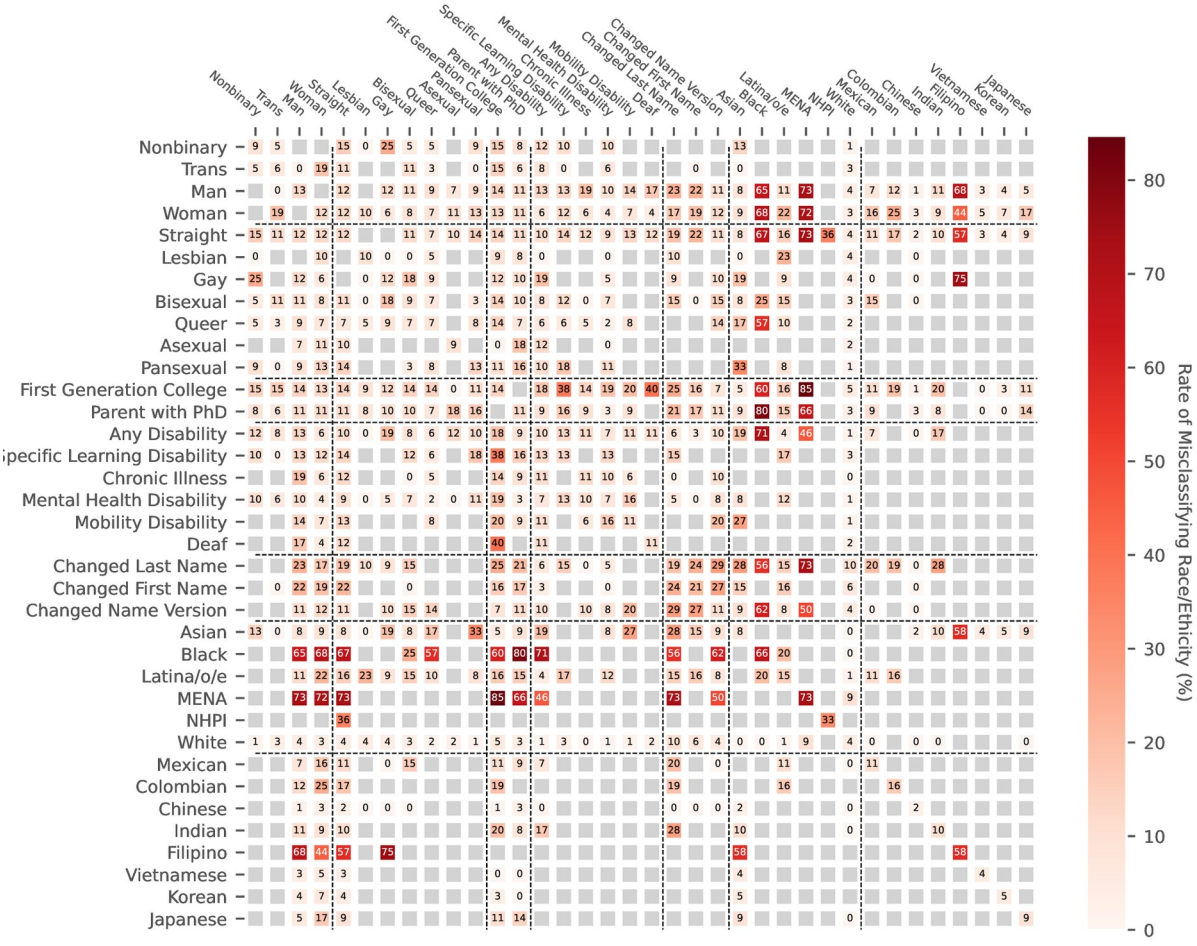

Figure 6: Two-way cross-tabulation of racial/ethnic misclassification from the `predictrace` algorithm on social science authors. Numbers are percentages. Cells with fewer than 10 people are grayed out and not reported.

Figure 6, a two-way cross-tabulation of race ascription error rates, reveals additional heterogeneity. For example, among Indian respondents, those whose parents did not go to college are more likely to be racially misclassified than those whose parent has a PhD (z=2.4, p=0.017, h=0.41), but the reverse is true for Black respondents: first generation scholars are *less* likely to be racially misclassified than those whose parent(s) have a PhD (z=2.4, p=0.015, h=0.46). Discussion

We have argued that cultural processes of naming and demographic membership interact in varied and complex ways, and we tested the relationships between demographic groups, names, and misrecognition. Here, we reflect on the heterogeniety observed above and its implications.

As others have noted, state-of-the-art name-based gender and race/ethnicity ascription algorithms are approaching the information-theoretic limit of accuracy beyond which additional reference data or more advanced modeling cannot improve performance.[13,47] Some names are low-information for a variety of reasons, including rare names, names commonly given to

multiple groups (e.g., men and women or Black and white Americans), and names where demographic correlations are lost in translation from their original writing/pronunciation to roman characters.

This has unequal effects across groups. For example, there is considerable heterogeneity in rates of misgendering within the category 'Asian.' Our results show that Chinese, Vietnamese, and Korean people are misgendered much more than Indian, Japanese, and other Asian-origin people. A naive machine learning impulse might be to gather more training data fornational origins that perform poorly. . That may work for Ethnicolor's North Carolina model, which underperforms its counterpart built on Florida data.

But this approach misses a more fundamental issue. English language publications romanize other languages by converting writing, including personal names, to Latin characters. English scientific databases like the Web of Science and computational researchers often go further, standardizing writing to a narrow subset of Latin characters with few or no diacritics, such as ASCII, for the sake of computational processing. For some languages, especially tonal languages this removes linguistic information that often carries demographic associations. Consider the following Mandarin example: 张伟 and 张薇 are both names, one masculine and the other feminine, but they both romanize to the same string: "Zhang, Wei," making it impossible to recover the original gender associations when only the romanized string is available.

Thus English name based gender imputation will always disproportionately misgender people from language groups where gender information might exist in naming, but is not carried over into English databases. While algorithms exist to impute gender from names written in Chinese and other languages, the increasing solidification of English as the global lingua franca of academic research[48] means that these problems are more a matter of the politics of language than technical challenges. Meanwhile, naming systems common in Spanish carry much more gender information than average into English databases and analyses. The increased information results in a reduction of misgendering. This comparatively better accuracy, however, poses the risk of overconfidence: users of these tools may forget or neglect that they still misgender people when working with Latina/o/e populations.

The unequal demographic information content of names that leads to heterogeneity in error rates is not only a language problem, but a sociocultural one. Due to the long history of slavery, there is considerable overlap between Black and white names in the US. The underrepresentation of Black people in most data sets means their race will be misrecognized more often than their white peers.[20] Moreover, within the US Black population, migration, social trends and movements, class, and other factors shape who goes by distinctively Black names, and thus who is ascribed Black identity by other people and algorithms.[17] Among the Black social scientists in our sample, those whose parent(s) have PhDs were correctly recognized as Black more often than those whose parent(s) did not attend college. This may be due to an interaction between education and race in how Black parents name their children. Or it may be due to an interaction between parental education and racialized names influencing which Black people are successful in academic careers. Whatever the process, it is not solely a function of class/parental education, because the pattern is reversed for Indian academics. Both the specific cultural context of naming, including race, national origin, education, and venue (e.g., author bylines in *American Sociological Review* differ from display names on Twitter), as well as the context of ascription (e.g. Who is inferring race? What is their reference population?) are critical

to understanding the racialization of names.

Separate from naming, the correlations among demographics in the social world can pose significant confounding challenges. The case of disability is instructive: trans and nonbinary people are much more likely than average to report disabilities, and also much more likely to be misgendered. Disabled people are also more likely to be misgendered. There are many plausible causal pathways for this relationship: disabled people may reflect more on their bodies and genders; trans and nonbinary people face adversity that may cause disability; gender transition often involves contact with psychologists, which could increase diagnosis for mental health disabilities. But even among cisgender people, those with disabilities are misgendered 60% more often ($p=0.001$). The area is ripe for qualitative analysis of gender, disability, and naming, informed by crip and trans theory. Analysis with name-ascription tools can help bring such associations to light, but it cannot account for them in the way other work can.

The problems of heterogeneous errors we identify generalize to other demographic factors, and can be exacerbated in unpredictable ways by attempts to reduce errors. Other work reveals errors in name-based racial ascription are heavily correlated with income, education, and census-tract level geography, especially when geography is used as a covariate to improve the overall accuracy of name-based inference.[26]

The heterogeneity in error rates with name-based demographic ascription can pose serious challenges to inference. For example, if Peng et al.[2] wanted to expand their analysis of discrimination against East-Asian people to also study discrimination against women, their analysis would likely be confounded by the fact that nearly half of Chinese women academics are incorrectly labeled as men. Attempts to correct for these inequalities in error rates can be thrown off by them. For example, Kozlowski et al.'s[20] approach to compensating for the high rate at which Black people are racially mislabeled assumes they are all mislabeled at the same rate. If their corrected data was used in an analysis of parental education or class, however, the uneven rates of racial misclassification for parental education would likely still confound their analysis. Further, no uniform adjustment can be made for parental education, as the direction of its effect is different for different subpopulations. The problem runs deep. And while these studies use academic authors as their target population, the demographic profile of their subjects is likely at least slightly different from the authors in our survey. To know the exact error profile in any particular application of these tools, one would need to repeat an analysis like ours in that specific context.

The high and highly heterogeneous error rates we demonstrate should give the many research, government, and corporate users of name-based demographic inference pause. Mislabeling people's gender, race/ethnicity, and other traits can have serious consequences, as discussed above. Moreover, errors can spill over in unexpected ways to create substantial biases in inferences about even seemingly unrelated groups, such as people with disabilities, Chinese women, or first-generation Black social scientists.

In light of this, we suggest five principles for conducting name-based demographic inference.

1. **Critical refusal**. Sometimes the right answer to "should we build or use this technology?" is simply "no."[49] Scholars and others are generally content not to infer sexuality, disability, class, and myriad other traits from names, even though that demographic information might be useful. Yet, it is common to infer gender, race, and ethnicity from names because many mistakenly believe that doing so is theoretically

justified, empirically effective, and ethically unproblematic. Those conditions are rarely met, which is why Mihaljević et al. conclude "Gender-inclusive bibliometric analyses can become possible only when no names or photographs are used as proxies for gender."[5] We would add that the same is true for race/ethnicity.

2. **Align the mechanism with the method.** Name-based demographic inference is a method that measures by external ascription, so studies concerned with external ascription are appropriate. Studies interested in self-identity, legal status, or biomarkers are not. For example, Peng et al.[2], evaluate whether authors with "East-Asian" names are discriminated against in the academic publication process compared to authors with "British-origin" names. Their proposed mechanism of discrimination and their measure of it are the same: ethnicity inferred from names. Similarly, Lagos[31] uses disagreement between voice-based gender inference and self-reported gender to construct a measure of misgendering, which enables important analyses of health disparities. Studies like these acknowledge that ascribed race and gender are important parts of race and gender experience, without confusing them for the whole truth or for individuals' sense of identity.
3. **Conduct inference specific to a population using domain expertise.** Jensen et al.[18] use their knowledge of the Indonesian regency of Indramayu, where the choice of Javanese, Indonesian (Bahasa), or Arabic names for children is a strong signal of religiosity, to develop a custom name-based religiosity inference model that works well in this setting, but would not translate to many other contexts. More generally, because demographic patterns change across populations by time, place, and other factors, imputation models will be more accurate when they are trained on the same population they are applied to.[13]
4. **Use subgroups with high accuracy.** Rather than attempting a universal model of racialization, Peng et al.[2] work only with groups that have high accuracy (East Asian and British origin names). Accuracy in differentiating white and Black Americans based on names is poor, and their use of the category 'British origin' instead of 'white' and 'Black' limits their analysis of name based discrimination to more supportable claims. This means not all research questions of substantive interest can be studied with these tools.
5. **Use only aggregate estimates of demographics from names, and check accuracy and bias on the target population**. Aggregate estimates, such as the percent of a population who are men, do not require individual ascriptions, and we can quantify their biases by surveying a subpopulation. For example, we might use our WoS data to compute the proportion of sociology authors who are men from their names. Because we conducted a survey, we know that the error rate in our specific population, when aggregated at the population level, is 4%. We further know it is biased to overcount men, undercount women, and exclude all nonbinary scholars. That information would allow us to compare the estimate of men's authorship in sociology with NSF data on PhDs granted or American Sociological Association data on membership. In contrast, if we used the imputed gender as a variable in regression, treating it as an individual predictor and not an aggregate summary, the systematic and highly variable misgendering of different subpopulations would create confounding with covariates like sexuality, disability, and race/ethnicity.

Developers of these tools can also learn from our results. For example, it may be responsible to only report aggregate statistics about input names, rather than individual predictions. Or, when presenting individual predictions, developers can help users appropriately apply and interpret their results by presenting data such as we present in this paper including information regarding variation in model accuracy across different groups. One common way developers have sought to increase overall accuracy is by adding covariates such as time and geography; however, recent research suggests this likely exacerbates error rate heterogeneity [26], making reporting especially important.

Further, developers could give users a clearer picture of the relationship between demographic characteristics and names by reporting two kinds of "unknowns," alongside their known category predictions: unknown unknowns (i.e., names for which little or no data exists), and known unknowns (i.e., names for which substantial data exists but demographic profiles are mixed). This distinction both provides users clarity about the demographics of names and respects people's choice of names that do not carry strong demographic signals. There is a robust literature in algorithmic fairness about designing algorithms in order to equalize error rates across groups, generally at the cost of overall performance, from which designers of demographic imputation tools might borrow. There is also a business case for optimizing these tools in the aforementioned ways, as users prefer tools that are more transparent and less biased. In turn, users should prioritize selecting tools that transparently report their performance across diverse subpopulations and tools that make an effort to minimize disparities across groups.

Important decisions about people's lives are increasingly made by computer algorithms. Governments, companies, and researchers deploy artificial intelligence algorithms in ways that can lead to unequal outcomes. From sorting résumés for job applications[50] to profiling social media users[12] to recommending sentence lengths and early parole for convicted offenders in the penal system[51,52], built-in biases in software systems shape our lives.[53] When data have incomplete or missing demographic data, there are incentives to fill these gaps with imputation. The resulting use of algorithms has important implications not only for how we perform and read science, but also for how we automate inequality.[54–56] Interrogating name-based demographic ascription is important for ensuring our methodologies are ethically responsible, empirically valid, and theoretically just.

Methods

Data

Using an institutional copy of the Web of Science (WoS) database, we selected all 139,882 unique email addresses for people who were listed as an author on an article in a sociology, economics, or communications journal (as defined by the Scimago Journal Rankings rankings) between 2016 and 2020. In compliance with relevant ethical regulations and with approval from the University of Connecticut IRB, we sent a link to each address asking authors to take a demographic survey with no compensation. Non-respondents received second and third follow-up reminders. In all, 19,924 people provided informed consent to take the survey. Responses from 16 people were discarded as unreliable because the respondents wrote things like "fuck you asshole," "this is woke bullshit," or "Apache Helicopter" in the open-ended self-identification questions. We believe the rate of hostile behavior was low because participation

was not anonymous. Our overall response rate was 14%. For this paper, we are interested in the correspondence between automated inference and self-reported demographics, rather than the generalizability of our sample to other populations. We note that each population is likely to differ in demographic profile, such that overall aggregate error rates may differ between populations, while the error rates we identify for demographic subgroups (e.g., Chinese women) are likely to be more robust.

Our survey asked a series of demographic questions available in the Supplemental Information. Importantly, we used two questions for gender: one for current gender with exclusive options for man, woman, nonbinary, and self-describe, and a separate yes/no question for whether the person considers themselves trans. Both gender questions were presented together. Note that nonbinary is not a subset of trans. In our sample, 53% of trans people are nonbinary, and 36% of nonbinary people are trans. Further, trans is not mutually exclusive with men or women.

Our race/ethnicity question used categories from the US Census and Pew Research, including national origin follow-up questions for people who selected Asian or Hispanic or Latina/o/e. Both the main and follow-up race/ethnicity questions had write-in options. Notably, many authors of English language social science publications live and work in places where the official US terms and categories of racial classification do not make as much sense. 2.9% of people chose not to answer the question, and 5.9% chose to write in an alternative description of themselves. We use the responses from the remaining 91% of authors who placed themselves into US administrative racial/ethnic categories, regardless of what country they work in. Similarly, the response options for parents' education followed the US education system, and some respondents chose not to use them. Whenever a participant skipped a question or wrote in an alternate answer, they were omitted from this paper's analysis of that question. As such, our results should be interpreted as holding among people who placed themselves inside the categories we name. The complete set of demographic questions is reprinted in the SI.

WoS provides display names from published English language articles in ASCII format. We parsed the names into given and surnames using the python package 'nameparser,' which handles a wide variety of linguistic/cultural naming conventions and written formats. Where given names were just initials, we use middle names as given names, unless those are also initials.

Demographic Ascription

We use four popular gender ascription algorithms: genderize.io, M3-Inference, R's `predictrace` package, and R's `gender` package.[23,57–59] Each relies on a different underlying corpus of names and method of inference (from simple dictionary lookup to neural networks). Similarly, we use four popular race/ethnicity ascription algorithms: ethnicolor's Florida voter model, ethnicolor's North Carolina voter model, the R package predictrace, and the R package wru.[59–61] A number of these models can incorporate additional information beyond names, such as age, country, location within the US, twitter biographies, or even a photograph to improve their predictions. Where WoS provided the country of the institution where an author is affiliated, we passed this information on to the algorithm that could use it (namely, genderize.io). The other information was not available in WoS and typically is not available in many applications for which name-based demographic imputation is used.

Errors

We labeled a gender classification as an error if an algorithm labeled someone 'man,' 'male,' or 'M' and they did not label themselves as a man in our survey, or if the algorithm labeled them 'woman,' 'female,' or 'F' and they did not label themselves as a woman in our survey. Most algorithms default to a 50% threshold for converting predicted probabilities to gender classifications. For algorithms that returned predicted probabilities, we used the 50% threshold. When an algorithm returned 'unknown' gender or a missing value in the probability vector, we omitted that data point from our analysis. This way we only evaluate algorithms on the data that they were confident enough to give predictions for. Some researchers arbitrarily set higher confidence thresholds. To ensure our results are robust and apply to those use cases also, we repeat our analysis using a 99% confidence threshold. The substantial heterogienety in error rates between demographic groups we show in the main analysis persists even when using this extreme threshold (Figure S3).

We took a conservative approach to labeling racial/ethnic classifications as errors, defining an error narrowly so that the tools would get the benefit of the doubt. If any of an algorithm's labels for a name matched any of the labels the person chose for themselves in the survey, we marked it correct. If an algorithm predicted "two or more races" and the person selected two or more, we marked it correct. And if an algorithm labeled someone "other" race and that person either labeled themselves "other" or they labeled themselves with a category that is not in the algorithm's repertoire (e.g. Native Hawaiian / Pacific Islander), we labeled it correct. We dropped cases where the algorithm did not make a prediction. If none of the race/ethnicities predicted by an algorithm match anything the respondent selected in the survey, or if the algorithm specified "non-Hispanic" and the person selected Hispanic or Latina/o/e, then we marked it as an error. Most algorithms offer a prediction that is the highest probability category or categories if several are equally likely. Where the algorithms offer only predicted probabilities, we do the same. `Predictrace` offers separate predictions for first and last names; we combined them such that each person's prediction was the union of all predictions for their given and surnames. Methodologies unable to stand up to our conservative test of the problem are inappropriate for most applied uses, where a stricter approach requiring exact matching (i.e., no extra or missing labels) is critical for mitigating racial misrecognition and for overall quality of inference.

Analysis

Most analyses are simple proportions of misrecognition, tabulated for different demographic subpopulations. This descriptive analysis demonstrates substantial heterogeniety and guides our theoretical discussion about some sources of that heterogeniety. In figures, we omit results for subgroups with fewer than 10 people, both because small group proportions are unreliable and to ensure k-anonymity of our respondents. When directly comparing groups in the text the text, we perform two-tailed z-tests of whether the proportion of errors differs between the groups and report effect sizes as Cohen's h value. These analyses are exploratory and descriptive, meant to bring to light a set of problems that are necessarily context dependent rather than to provide confirmatory point estimates of invariant quanitites or causal explanations of underlying relationships.

Data availability

Web of Science data are available from Clarivate Analytics, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The survey data that support the findings of this study are not publicly available because they contain information that could compromise research participant privacy/consent. Non-identifying aggregate data are available upon reasonable request to the corresponding author, JL. "Reasonable requests" should come from researchers with an active institutional affiliation, be for research purposes only, and have ethical approval from their Institutional Review Board or appropriate oversight body. Requests would be subject to a data sharing agreement. The authors commit to maintaining the raw data associated with this study for a minimum of five years. Source data for all figures is available with the supplimental materials in an Open Science Framework repository: https://osf.io/avzpk.

Code Availability

While the results we present are simple statistics, code to generate our results and figures is available with the supplimental materials in an Open Science Framework repository: https://osf.io/avzpk.

References

1. Matias, J. N., Szalavitz, S. & Zuckerman, E. FollowBias: Supporting Behavior Change toward Gender Equality by Networked Gatekeepers on Social Media. in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* 1082–1095 (Association for Computing Machinery, 2017). doi:10.1145/2998181.2998287.
2. Peng, H., Lakhani, K. & Teplitskiy, M. Acceptance in Top Journals Shows Large Disparities across Name-inferred Ethnicities. Preprint at https://doi.org/10.31235/osf.io/mjbxg (2021).
3. Hofstra, B. & de Schipper, N. C. Predicting ethnicity with first names in online social media networks. *Big Data Soc.* **5**, 2053951718761141 (2018).
4. King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J. & West, J. D. Men Set Their Own Cites High: Gender and Self-citation across Fields and over Time. *Socius* **3**, 2378023117738903 (2017).
5. Mihaljević, H., Tullney, M., Santamaría, L. & Steinfeldt, C. Reflections on Gender Analyses of Bibliographic Corpora. *Front. Big Data* **2**, (2019).
6. Keyes, O. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.* **2**, 1–22 (2018).
7. D'Ignazio, C. A Primer on Non-Binary Gender and Big Data. *MIT Center for Civic Media* https://civic.mit.edu/index.html%3Fp=1165.html (2016).
8. Lockhart, J. W. Gender, Sex, and the Constraints of Machine Learning Methods. in *Oxford Handbook of the Sociology of Machine Learning* (eds. Borch, C. & Pardo-Gurrera, J. P.) (OUP, 2023).
9. Santamaría, L. & Mihaljević, H. Comparison and benchmark of name-to-gender inference services. *PeerJ Comput. Sci.* **4**, e156 (2018).
10. Lindsay, J. & Dempsey, D. First names and social distinction: Middle-class naming practices in Australia. *J. Sociol.* **53**, 577–591 (2017).

11. Bertrand, M. & Mullainathan, S. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *Am. Econ. Rev.* **94**, 991–1013 (2004).
12. Fosch-Villaronga, E., Poulsen, A., Søraa, R. A. & Custers, B. H. M. A little bird told me your gender: Gender inferences in social media. *Inf. Process. Manag.* **58**, 102541 (2021).
13. Van Buskirk, I., Clauset, A. & Larremore, D. B. An Open-Source Cultural Consensus Approach to Name-Based Gender Classification. Preprint at http://arxiv.org/abs/2208.01714 (2022).
14. West, C. & Zimmerman, D. H. Doing Gender. *Gend. Soc.* **1**, 125–151 (1987).
15. Bonilla-Silva, E. The Essential Social Fact of Race. *Am. Sociol. Rev.* **64**, 899–906 (1999).
16. Seguin, C., Julien, C. & Zhang, Y. The stability of androgynous names: Dynamics of gendered naming practices in the United States 1880–2016. *Poetics* **85**, 101501 (2021).
17. Fryer, R. G., Jr. & Levitt, S. D. The Causes and Consequences of Distinctively Black Names*. *Q. J. Econ.* **119**, 767–805 (2004).
18. Jensen, J. L. *et al.* Language Models in Sociological Research: An Application to Classifying Large Administrative Data and Measuring Religiosity. *Sociol. Methodol.* **52**, 30–52 (2022).
19. Lieberson, S., Dumais, S. & Baumann, S. The Instability of Androgynous Names: The Symbolic Maintenance of Gender Boundaries. *Am. J. Sociol.* **105**, 1249–1287 (2000).
20. Kozlowski, D. *et al.* Avoiding bias when inferring race using name-based approaches. *PLOS ONE* **17**, e0264270 (2022).
21. Sebo, P. Using genderize.io to infer the gender of first names: how to improve the accuracy of the inference. *J. Med. Libr. Assoc.* **109**, 609–612 (2021).
22. Müller, D., Te, Y.-F. & Jain, P. Improving data quality through high precision gender categorization. in *2017 IEEE International Conference on Big Data (Big Data)* 2628–2636 (2017). doi:10.1109/BigData.2017.8258223.
23. Wang, Z. *et al.* Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. in *The World Wide Web Conference* 2056–2067 (Association for Computing Machinery, 2019). doi:10.1145/3308558.3313684.
24. Silva, G. C., Trivedi, A. N. & Gutman, R. Developing and evaluating methods to impute race/ethnicity in an incomplete dataset. *Health Serv. Outcomes Res. Methodol.* **19**, 175–195 (2019).
25. Mateos, P. A review of name-based ethnicity classification methods and their potential in population studies. *Popul. Space Place* **13**, 243–263 (2007).
26. Argyle, L. & Barber, M. Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records. *Am. Polit. Sci. Rev.* (2023).
27. ASA. ASA Membership. *American Sociological Association* https://www.asanet.org/academic-professional-resources/data-about-discipline/asa-membership (2021).
28. Kessler, S. J. & McKenna, W. *Gender: An Ethnomethodological Approach*. (University of Chicago Press, 1985).
29. Pascoe, C. J. *Dude, You're a Fag: Masculinity and Sexuality in High School*. (University of California Press, 2007).
30. McNamarah, C. T. Misgendering. *Calif. Law Rev.* **109**, 2227–2322 (2021).
31. Lagos, D. Hearing Gender: Voice-Based Gender Classification Processes and Transgender Health Inequality. *Am. Sociol. Rev.* **84**, 801–827 (2019).

32. Browne, K. Genderism and the Bathroom Problem: (re)materialising sexed sites, (re)creating sexed bodies. *Gend. Place Cult.* **11**, 331–346 (2004).
33. Whitley, C. T., Nordmarken, S., Kolysh, S. & Goldstein-Kral, J. I've Been Misgendered So Many Times: Comparing the Experiences of Chronic Misgendering among Transgender Graduate Students in the Social and Natural Sciences. *Sociol. Inq.* **n/a**, (2022).
34. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. 10 (1979).
35. Hamidi, F., Scheuerman, M. K. & Branham, S. M. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* 8:1-8:13 (ACM, 2018). doi:10.1145/3173574.3173582.
36. Scheuerman, M. K., Pape, M. & Hanna, A. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data Soc.* **8**, 20539517211053710 (2021).
37. Bourg, C. Gender mistakes and inequality. (Stanford University, 2003).
38. Davis, G. & Preves, S. Intersex and the Social Construction of Sex. *Contexts* **16**, 80–80 (2017).
39. Fausto-Sterling, A. *Sexing the body: gender politics and the construction of sexuality*. (Basic Books, 2000).
40. Lockhart, J. W. Paradigms of Sex Research and Women in STEM. *Gend. Soc.* **35**, 449–475 (2021).
41. Science must respect the dignity and rights of all humans. *Nat. Hum. Behav.* **6**, 1029–1031 (2022).
42. Slater, R. B. The Blacks who First Entered the World of White Higher Education. *J. Blacks High. Educ.* 47–56 (1994) doi:10.2307/2963372.
43. Blumenfeld, W. J. On The Discursive Construction Of Jewish "Racialization" And "Race Passing:" Jews As "U-boats" With A Mysterious "Queer Light". *J. Crit. Thought Prax.* **1**, 2 (2012).
44. Nakamura, L. Cyberrace. *PMLA* **123**, 1673–1682 (2008).
45. Sims, J. P. Reevaluation of the Influence of Appearance and Reflected Appraisals for Mixed-Race Identity: The Role of Consistent Inconsistent Racial Perception. *Sociol. Race Ethn.* **2**, 569–583 (2016).
46. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Machine Learning Research* 1–15 (2018).
47. Tzioumis, K. Demographic aspects of first names. *Sci. Data* **5**, 180025 (2018).
48. Di Bitetti, M. S. & Ferreras, J. A. Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications. *Ambio* **46**, 121–127 (2017).
49. Garcia, P. *et al.* No: Critical Refusal as Feminist Data Practice. in *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* 199–202 (Association for Computing Machinery, 2020). doi:10.1145/3406865.3419014.
50. Caplan, R., Donovan, J., hanson, L. & Matthews, J. *Algorithmic Accountability: a Primer*. https://datasociety.net/wp-content/uploads/2019/09/DandS_Algorithmic_Accountability.pdf (2018).
51. Angwin, J., Larsen, J., Mattu, S. & Kirchner, L. *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

(2016).
52. Harcourt, B. E. Risk as a Proxy for Race: The Dangers of Risk Assessment. *Fed. Sentencing Report.* **27**, 237–243 (2015).
53. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
54. Eubanks, V. *Automating inequality: how high-tech tools profile, police, and punish the poor*. (St. Martin's Press, 2017).
55. O'Neil, C. *Weapons of math destruction: how big data increases inequality and threatens democracy*. (Crown, 2016).
56. Benjamin, R. *Race After Technology: Abolitionist Tools for the New Jim Code*. (Polity, 2019).
57. Genderize.io | Determine the gender of a name. https://genderize.io/.
58. Mullen, L., Blevins, C. & Schmidt, B. gender: Predict Gender from Names Using Historical Data. (2021).
59. Kaplan, J. predictrace: Predict the Race and Gender of a Given Name Using Census and Social Security Administration Data. (2021).
60. appeler/ethnicolor. (2022).
61. Khanna, K., Bertelsen, B., Olivella, S., Rosenman, E. & Imai, K. wru: Who are You? Bayesian Prediction of Racial Category Using Surname, First Name, Middle Name, and Geolocation. (2022).

Acknowledgements

The authors thank Michael Thompson-Brusstar for his insights. Grace Azzara, Gabriel Cash, Jocelyn Anaya Galvan, Kelly Lelapinyokul, Samantha Martinez, and Brooke Rose provided excellent research assistance. The authors received no funding specifically for this work. Financial support for research assistants was in part provided by a College of Arts and Sciences Dean's Grant to MMK from Santa Clara University.

Author Contributions

JWL designed and executed the analyses. JWL and MMK wrote the manuscript. All authors contributed to designing the survey and revising the manuscript.

Competing Interests

The authors declare no competing interests.