Beyond Appearance: The Socioeconomic and Historical Roots of Racial Identity in Brazil

Diogo Baerlocher
University of South Florida

Renata Caldas
University of South Florida

Francisco Cavalcanti
UFPE

August, 2025*

Abstract

Racial identity is not solely a matter of physical appearance but is also shaped by social and historical context. Using data on over 500,000 candidates for local office in Brazil's 2020 elections, we study how self-reported race—specifically, identification as *white*—relates to phenotypic appearance, socioeconomic characteristics, and local social perceptions. We use machine learning to extract appearance-based probabilities of racial classification from candidate photographs and show that these probabilities explain a significant share of variation in self-reported race. Socioeconomic factors such as education, gender, and wealth also influence racial identification, though their effects diminish among individuals whose appearance more clearly aligns with the *white* category. Municipality fixed effects, which we interpret as capturing local social perception bias, vary systematically across regions and are strongly associated with historical slave population shares. We further show that areas with state-sponsored European settlements—often associated with more inclusive institutions—exhibit lower rates of *white* self-identification, contrasting with the positive association between slavery intensity and *white* identification. Our findings highlight the enduring role of social and historical forces in shaping racial classification and suggest that racial inequality cannot be fully understood without accounting for the social construction of race.

Keywords: Racial Classification, Social Identity, Phenotypic Appearance, Historical Legacy
JEL Codes: J15, N36, Z13

**Baerlocher:* Department of Economics, University of South Florida, 4202 E. Fowler Ave. CMC342, Tampa, FL, 33620. (email: baerlocher@usf.edu). *Caldas:* Department of Economics, University of South Florida, 4202 E. Fowler Ave. CMC342, Tampa, FL, 33620. (email: renatad@usf.edu). *Cavalcanti:* Department of Economics, Universidade Federal de Pernambuco (UFPE), Avenida dos Economistas, s/n, Centro de Ciências Sociais Aplicadas/UFPE, Cidade Universitária Pernambuco, Brasil CEP 50740-580. (email: francisco.dlcavalcanti@ufpe.br). We are thankful to Linh Tran for excellent research assistance.

1 Introduction

Racial inequality remains a pervasive feature of modern societies, influencing access to education, employment, income, political power, and justice (Derenoncourt, Kim, Kuhn and Schularick, 2024; Lang and Spitzer, 2020). Understanding the roots and persistence of these disparities requires careful attention not only to outcomes but also to how racial categories themselves are constructed and maintained. Current scholarship increasingly recognizes that racial classification is not only biologically meaningless (Morning, 2011), but also deeply embedded in social norms, cultural narratives, and historical legacies (Omi and Winant, 2014). How individuals are perceived racially—and how they identify themselves—reflects these layered social processes, making racial identity both fluid and context-dependent. This complexity is essential to understanding how race contributes to persistent disparities in economic and social outcomes.

This paper investigates how self-reported racial identity in Brazil is shaped by physical appearance, socioeconomic characteristics, and social context. Brazil is marked by significant racial diversity, the result of European colonization and the slave labor of millions of Africans. This historical mix, along with the presence of Indigenous populations, has produced a wide range of phenotypic traits that do not always align neatly with broad racial categories such as *white* or *black*. Although the mixed-race category *pardo* is officially recognized in national statistics, racial identity in Brazil is widely understood to be a fluid and socially constructed phenomenon (Telles, 2004).

A key challenge in studying racial self-identification is obtaining comparable measures of individuals' physical appearance, which are essential for analyzing how people with similar phenotypic traits identify racially. To address this, we use a dataset consisting of Brazilian candidates for local public office in the 2020 municipal elections. As part of Brazil's electronic voting system, the Electoral Court collects and publicly releases candidate photographs, which must meet specific quality standards. We apply machine learning techniques to these images to estimate, for each individual, the probability of being classified into one of six broad ethnoracial categories: *Asian*, *black*, *Indian*, *Latino/Hispanic*, *Middle Eastern*, and *white*. These probabilities serve as a multidimensional measure of each candidate's appearance. This approach yields a nationwide sample with phenotypic information for over 500,000 individuals.

Importantly, candidates are also required to report a range of socioeconomic characteristics, including self-identified race, gender, age, marital status, occupation, and educational attainment. We use self-reported race as our main dependent variable, constructing a binary indicator equal to one if the candidate identified as *white* and zero otherwise. The remaining variables serve as controls to explain variation in self-identification. One important characteristic not reported by candidates is income. As a proxy, we use campaign spending per capita. Additionally, because we focus on local elections, each candidate can be directly linked to a specific municipality. We leverage this structure to include municipality fixed effects, which allow us both to control for and quantify local-level variation in social perceptions of race.

Our results suggest that variation in phenotypic appearance explains approximately 22% of the variation in self-reported race. Compared to the reference category, *Latino/Hispanic*, individuals with a higher probability of being classified as *white* or *Middle Eastern* are more likely to self-identify as *white*, while those with a higher probability of being classified as *Indian* or *black* are less likely to do so. The effect of a higher probability of being classified as *Asian* is relatively small compared to *Latino/Hispanic*. These findings align with expectations and support the use of machine learning-based probabilities as a valid proxy for phenotypic appearance.

With respect to socioeconomic characteristics, we find that the marginal effects of all variables on self-reported race are influenced by individuals' phenotypic appearance. Once we control for machine learning-based appearance probabilities, the estimated marginal effects decrease substantially. For example, individuals with a college degree are about 15 percentage points more likely to self-identify as *white* than those with less than primary education; however, this difference drops to approximately 8 percentage points when controlling for appearance. Importantly, the effects remain statistically significant even after accounting for appearance. Men are more likely than women to self-identify as *white*, although this result is not robust to the choice of face-detection algorithm. Married and single individuals are less likely to do so compared to those who were previously married. Higher education, older age, and greater wealth are also positively associated with identifying as *white*. These results highlight how social and economic status can shape an individual's perception of their racial identity.

Additionally, to examine whether these effects are driven by ambiguous appearance relative to broad racial categories, we compute the marginal effects of socioeconomic characteristics conditional on the machine learning algorithm's estimated probability of being classified as *white*. The results show that as the predicted probability of being *white* increases, the marginal effects of socioeconomic variables decline substantially. This suggests that when an individual's appearance clearly aligns with the *white* category, socioeconomic characteristics play a smaller role in shaping racial self-identification.

Finally, we investigate the institutional and historical roots of social perceptions of race. Specifically, we explore how local social norms shape the way individuals self-identify racially. To capture this dimension, we estimate municipality fixed effects that hold constant individuals' observable characteristics and thus reflect the local propensity to classify someone as *white*, conditional on appearance and socioeconomic status. We interpret this estimated component as a measure of social racial perception bias. This bias is significantly higher in municipalities located in the Southeast and South of Brazil. We hypothesize that this pattern is linked to distinct historical trajectories in these regions, particularly the legacy of slavery. Areas with a larger enslaved population during the colonial period likely developed stronger social hierarchies and more rigid racial boundaries, which may have persisted over time as enduring local norms of racial classification. Consistent with this interpretation, we find that the share of enslaved individuals in 1872 is positively associated with higher levels of social racial perception bias, even after controlling for state fixed effects and geographic characteristics.

To extend this analysis, we ask whether the results above are driven by enduring social beliefs about racial hierarchy — likely stronger in areas with more intensive slavery — or whether they reflect different colonization patterns that produced distinct institutions in later years. We explore these channels by estimating the effect on racial self-classification of having received state-sponsored European settlements in the early twentieth century, following the approach of Rocha, Ferraz and Soares (2017). Since Europeans were recruited by the state to replace what was considered inferior slave labor, these settlements may have fostered and perpetuated notions of racial inferiority. On the other hand, it is well known that European settlements tended to introduce more inclusive institutions, in contrast to the extractive institutions often associated with slavery-based colonization. Our results show that state-sponsored European settlements have a negative effect on the probability of self-identifying as *white*, in contrast to the positive effect of slavery intensity. These findings suggest that the extractive institutions fostered by slavery contributed to a persistent social bias toward *white* self-classification, as they rooted racial hierarchies and linked social status to whiteness.

Related Literature. Our paper contributes to a growing literature on the determinants of racial identity. This body of work highlights the limitations of relying solely on self-identified race, especially in contexts where racial classification is fluid, context-dependent, and shaped by social, regional, and economic factors (Bailey and Telles, 2006; Telles and Paschel, 2014). In a recent review, Saperstein (2025) documents a high degree of temporal fluidity in racial identity, particularly among individuals who initially identified as *non-Hispanic Asian* or *Hispanic*. The author concludes that "Where studies [...] converge is in finding that when people change their racial or ethnic responses, they tend to do so in line with existing racial stereotypes and patterns of inequality" (p. 526). In many cases, individuals adjust their racial self-identification in response to their socioeconomic status, aspirations for upward mobility, or how they are perceived by others—rather than based solely on their personal sense of identity (Mitchell, 2010; Telles and Paschel, 2014). Examples range from studies showing that highly educated nonwhite parents are more likely to classify their children as white compared to their less-educated counterparts (Schwartzman, 2007), to evidence from Agadjanian and Lacy (2021) linking shifts from nonwhite to white racial identification with political realignment, particularly among new Republican supporters in 2016.

We contribute to this literature by providing new evidence on the socioeconomic determinants of racial identity in the Brazilian context. Unlike previous studies, we leverage computer vision tools to quantify individuals' phenotypic features, such as skin tone and facial morphology, as perceived by others. This approach differs from classifying individuals into discrete racial categories: rather than assigning a race, we measure the degree to which observable traits align with social perceptions of racial appearance. This allows us to compare self-identified race with perceived phenotype in a continuous and more nuanced way. de Lucena Coelho, Estevan, Nakaguma and Rabelo (2024) also use candidates' photographs in their analysis of how a mayor's race affects the racial composition of public employment. However, they employ machine learning algorithms to categorize candidates by race, whereas our method aims to measure variation in phenotypical characteristics. In this respect, our paper is closely related to Francis and Tannuri-Pianto (2013), who examine the effects of socioeconomic characteristics and racial quotas on the racial identification of Brazilian undergraduate students. Their analysis also uses student photographs to measure skin tone. We echo their findings by showing that males are more likely to self-identify as *white*, while also exploring additional socioeconomic characteristics—such as marital status, education level, and spatial heterogeneity—that their study does not address, given their smaller sample.

In contrast to Francis and Tannuri-Pianto (2013) and others (Antman and Duncan, 2023, 2024; Francis-Tan and Tannuri-Pianto, 2024), our focus is not on affirmative action policies but rather on the historical roots of social perceptions of racial identity. Specifically, we examine the legacy of African slavery during the colonial and imperial eras in shaping these social perceptions in Brazil. In doing so, our paper contributes to the broader literature on the long-term effects of the slave trade. A seminal contribution in this field is Nunn and Wantchekon (2011), who links exposure to slave raids during the slave trade period with contemporary levels of mistrust. We build on this tradition by connecting the historical intensity of slavery with present-day cultural norms related to racial identity. Our work also relates to studies examining the long-term effects of different forms of colonization. In the Brazilian context, Rocha et al. (2017) show that state-sponsored European settlements fostered enduring investments in education, while Naritomi, Soares and Assunção (2012) demonstrate that current institutional differences arose from colonial patterns shaped by successive sugar, gold, and coffee booms. Additionally, Laudares and Valencia Caicedo (2023) find that historical slavery in Brazil is associated with higher income inequality and a wider racial income gap today.

2 Historical and Sociopolitical Background

Brazil was colonized primarily by the Portuguese, although other European groups—such as Italians, Germans, French, and Dutch—also contributed to its settlement. Long before colonization, however, the territory was home to a diverse array of Indigenous peoples, whose cultures and populations were profoundly impacted by European conquest, disease, and displacement. Over more than three centuries of colonial rule, Brazil also became a major destination in the transatlantic slave trade, with millions of Africans forcibly brought to work on sugarcane plantations and in other forms of coerced labor. The interaction among Indigenous, African, and European populations led to widespread racial mixing, which remains a defining feature of Brazilian society.

Given these diverse racial and ethnic origins that shaped Brazilian society, some researchers have argued that race is a meaningless or nonexistent category in Brazil—a notion often referred to as the "racial democracy" myth. However, this view is contradicted by a wide body of empirical evidence showing persistent racial inequalities (Andrews, 1996; Htun, 2004; Dupree-Wilson, 2021). Data from the Brazilian Institute of Geography and Statistics (IBGE) indicate that Afro-descendants in Brazil continue to face structural disadvantages: they have lower average incomes, higher unemployment rates, lower educational attainment, and are disproportionately represented in informal labor markets (IBGE, 2024). They are also more likely to live in underserved areas with limited access to quality healthcare, education, and public services. These disparities highlight how race continues to operate as a powerful axis of social and economic inequality in contemporary Brazil.

The Brazilian Census currently uses the term *pardo* to encompass all people of mixed racial backgrounds, and the official statistics are based solely on self-classification. One of the primary challenges in analyzing racial inequalities using self-classified racial data is the fluid and socially constructed nature of racial identity. In Brazil, where racial boundaries are often ambiguous and heavily influenced by social, regional, and economic factors, individuals may self-identify in ways that do not align with how others perceive them. This subjective element can obscure patterns of discrimination that are based on phenotypic traits such as skin color, facial features, or hair texture—characteristics that may drive discriminatory behavior. As a result, relying solely on self-classification may underestimate the extent of racial inequality, particularly if individuals who phenotypically appear *black* or mixed-race identify as *white* for social or strategic reasons. Moreover, self-identification can be influenced by socio-economic mobility, with some individuals "whitening" their racial identity as they gain education or income, further complicating efforts to track structural racial inequalities through self-reported data alone.

Within this institutional context of fluid racial categories, broadly defined as *pardo*, individuals in Brazil self-identify their race in various settings, including Census questionnaires, college applications, and, relevant to our study, political candidacies. It is important to note that in 2020, the year of our study, there was no affirmative action policy formally benefiting candidates of any local race. Therefore, their racial declarations likely reflect their own notions of race, uninfluenced by the political competition.

3 Empirical Framework

3.1 Conceptual issues for racial self-identification

We hypothesize that an individual's self-reported race depends on three sets of variables: the phenotypic appearance of individual i living in community c, denoted by $z_{i,c}$; their socio-demographic characteristics, denoted by $x_{i,c}$; and the influence of their community, denoted by w_c. Let the reported race be represented by the binary variable $y_{i,c}$, indicating whether individual i in community c reported being *white*. We model this as

$$y_{i,c} = f(z_{i,c}, x_{i,c}, w_c) + \varepsilon_{i,c}, \tag{1}$$

where $\varepsilon_{i,c}$ is an error term assumed to have mean zero.

Our goal is to measure the relative importance of each set of variables in shaping individuals' self-reported race. If race were purely a matter of physical appearance, then $x_{i,c}$ and w_c would have no effect. Thus, identifying the role of phenotypic appearance is central to our analysis. In this context, we treat Brazilian municipalities as the relevant communities.

3.2 Data and Sample Construction

Our sample includes the universe of all candidates who ran for mayoral and city council offices in the 2020 Brazilian municipal elections—a total of 558,226 individuals. We focus on candidates in local elections for several reasons. First, they provide a large and diverse sample, much broader than using candidates in state or federal contests. Second, local candidates—especially those running for city council positions—are less likely to be "professional" politicians, which enhances the external validity of our findings. Third, local elections allow for an accurate linkage between each candidate and their municipality, enabling us to capture community-level effects.

A key advantage of using electoral candidates is the public availability of standardized photographs. Due to Brazil's electronic voting system, candidate photos are submitted by political parties when registering candidacies and must meet uniform formatting standards established by the Electoral Court.[1]

3.3 Measuring Phenotypic Appearance

We employ a facial recognition framework based on deep convolutional neural networks to infer the racial and ethnic appearance of each candidate from their photographs. The method follows a computer-vision pipeline that detects, aligns, and normalizes faces before classifying them into six broad racial categories: *Asian*, *Indian*, *black*, *white*, *Middle Eastern*, and *Latino/Hispanic* (Karkkainen and Joo, 2021; Serengil and Ozpinar, 2024). These algorithms, trained on large-scale datasets containing over 100,000 human faces, achieve accuracy levels comparable to or even surpassing human performance in facial recognition tasks. Rather than assigning a fixed label, the model produces a probability distribution over the six categories, thereby capturing the degree of uncertainty in racial appearance. For instance, the algorithm assigns U.S. President Donald Trump a 99.98% probability of being classified as *white*, while former U.S. President Barack Obama receives a 61.83% probability of being classified as *black* and a 21.40% probability of being

[1] Resolution No. 23,609 of December 19, 2019 (Art. 27, §2) specifies that photographs must have 161×225 pixel dimensions, 24-bit color depth, a uniform background, and depict the candidate in a frontal pose with appropriate attire. Ethnic or religious clothing and necessary accessories for persons with disabilities are permitted, while decorative or campaign-related elements are prohibited.

classified as *Latino/Hispanic*.[2][3]

The method is not without limitations. The performance of the algorithm depends on factors such as photo quality, lighting conditions, and the subject's pose. Additionally, the racial categories used by the algorithm are broad and based on U.S.-centric classifications, which do not fully capture Brazil's racial context, particularly the *pardo* category, which has no direct equivalent. Nonetheless, given that the photographs are submitted to the Brazilian Electoral Court and must meet minimum quality standards, we expect the algorithm to perform reasonably well. Moreover, for our purposes, the broad racial categories are not a drawback. We do not aim to definitively classify individuals' race but rather to capture their phenotypic appearance. Using the full set of classification probabilities enables us to construct a multi-dimensional measure of how individuals are likely perceived based on appearance. That is, individuals with similar sets of probabilities are likely to share similar physical traits.

3.4 Socioeconomic Characteristics

The set of individual socioeconomic variables is obtained from the candidacy information from TSE. Candidates report their race, gender, marital status, occupation, education level, and age. The reported race serves as our outcome variable. Although candidates may identify as *white*, *black*, *yellow*, *pardo*, or *Indigenous*, we construct a binary variable indicating whether the candidate reported being *white* (1) or *nonwhite* (0). All other variables are included in the vector $x_{i,c}$ of socioeconomic characteristics. To approximate individual wealth, we also include campaign spending per capita. The conclusions remain unchanged when using asset values as a measure of wealth (see Appendix B). In most estimations, we select *Latino/Hispanic* as the reference category for race classification. *Female* is the reference for gender, and *Less than primary education* for education. For marital status, we group *Separated*, *Divorced*, and *Widowed* individuals as *Previously Married*, which serves as the reference category.

[2] We implement this procedure using the *DeepFace* library in Python, which integrates several state-of-the-art facial recognition models (VGG-Face, FaceNet, ArcFace, among others) and performs all standard stages of modern facial analysis—detection, alignment, normalization, and classification. We rely on the default *opencv* detector for face detection, and confirm in Appendix C that results are robust when using the *RetinaFace* algorithm, with the notable exception of the sex coefficient discussed in the next section.

[3] Table A.2 shows a list of 32 influential individuals and their race classifications according to *DeepFace*. The list includes individuals who appeared four or more times in Time magazine's annual Time 100 list of most influential people. The photographs are taken from the Time 100 Wikipedia page (`https://en.wikipedia.org/wiki/Time_100#The_25_Most_Influential_People_on_the_Internet`); if the algorithm does not recognize a face, we use the image from the individual's main Wikipedia page.

3.5 Historical and Other Controls

To capture the historical legacy of slavery, we use data from the 1872 Brazilian Census, which provides the Share of Enslaved Population (1872), the ratio of enslaved individuals to the total population in each minimum comparable area of 1872. This variable is matched to modern municipalities using historical boundary correspondence. Geographic controls include the median temperature, median elevation, median rainfall, median terrain ruggedness index, log distance to rivers, log distance to the coast, and log area size of each municipality. Municipality-level controls, derived from the 2022 Population Census (IBGE), include the share of white individuals, share of males, share of literate individuals, and the nonwhite-to-white wage gap.

4 Racial Identity and Individuals' Characteristics

In Table 1, we present results from four regression models examining how individuals' phenotypical appearance, as inferred from their facial features, relates to their self-identified racial classification. The dependent variable is a binary indicator equal to one if the individual reported being *white*. In column (2), we estimate the model $y_{i,c} = \alpha + \beta_1 White_{i,c} + \beta_2 Asian_{i,c} + \beta_3 Black_{i,c} + \beta_4 Indian_{i,c} + \beta_5 MiddelEastern_{i,c} + \mu_c + \epsilon_{i,c}$, where $y_{i,c}$ denotes whether individual i in municipality c self-identifies as white, the regressors represent the algorithm-assigned probabilities that the same individual is classified as belonging to each racial category, and μ_c captures unobserved municipality-specific factors such as local norms or regional patterns of racial identification. The results show that the probability of being algorithmically classified as white is a strong predictor of self-identifying as white: the coefficient of 0.404 implies that a 10-percentage-point increase in the predicted probability of being classified as white is associated with a 4.0-percentage-point increase in the likelihood of self-identifying as white. The probability of self-identifying as white also rises with the algorithmic probability of being classified as Middle Eastern, while it decreases for all other categories, particularly for individuals classified as Black or Indian. Importantly, after accounting for municipality fixed effects and estimating a logistic specification in columns (3)–(4), the overall pattern of results remains robust.

The results confirm that phenotypical appearance, as detected algorithmically, plays a central role in racial self-identification and reveals systematic discrepancies between appearance and self-classification for certain groups. The positive association between a *Middle Eastern* appearance and self-identifying as *white* suggests that social perceptions in Brazil tend to align Middle Eastern features more closely with whiteness.

Table 1: **Racial Identify and Phenotypical Appearance**

	OLS		Logit	
	(1)	(2)	(3)	(4)
	Dependent Variable: Reported White			
Phenotypical classification				
White	0.617***	0.404***	2.49***	0.373***
	(0.005)	(0.004)	(0.026)	(0.004)
Asian	-0.027***	-0.007	0.144***	0.022***
	(0.006)	(0.005)	(0.028)	(0.004)
Black	-0.346***	-0.406***	-4.33***	-0.649***
	(0.006)	(0.008)	(0.070)	(0.010)
Indian	-0.622***	-0.526***	-2.90***	-0.434***
	(0.012)	(0.010)	(0.074)	(0.011)
Middle Eastern	0.329***	0.240***	1.18***	0.176***
	(0.005)	(0.004)	(0.024)	(0.003)
Municipality Fixed Effect		✓	✓	✓
Average Marginal Effect	✓	✓		✓
Observations	535,832	535,832	530,901	530,901
(Pseudo) R^2	0.21999	0.38932	0.33807	0.33807

Standard errors clustered at the municipality level are reported in parentheses. The dependent variable is a binary indicator equal to 1 if the individual self-reported as white. Regressors are the algorithm-assigned probabilities that the individual is identified as each respective race. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In contrast, the negative coefficients for *black* and *Indian* appearances, where the latter may be more easily associated with the mixed race or *pardo* category, are consistent with expected patterns. The relatively small coefficient for *Asian* appearance is notable: it reflects that, in general, individuals of Asian descent often self-identify as *white* but may also report as *yellow*. The R^2 in column (1) is 0.22, indicating that variation in phenotypical appearance explains 22 percent of the variation in self-reported race. While substantial, this also implies that other factors play an important role. In column (2), the inclusion of municipality fixed effects increases the R^2 to 0.39, suggesting that local social perceptions are as influential as physical appearance in shaping racial identity.

We next examine the effect of socioeconomic characteristics on racial self-identification. It is well documented that individuals who identify as *white* tend to achieve higher levels of education and income, largely because racial identification correlates with historical and structural differences in access to schooling and labor market opportunities. We therefore ask whether individuals with more favorable socioeconomic characteristics are also more likely to self-identify as *white*, conditional on their phenotypical appearance and local social context. Figure 1 presents the marginal effects of various socioeconomic variables on the likelihood of reporting as *white*, estimated from a logistic regression model with municipality fixed effects. For each variable, the figure displays two estimates: one that does not control for phenotypical appearance and one that does. This comparison allows us to assess the extent to which socioeconomic influences on racial identity operate directly or through correlations with appearance.

Our findings indicate that men are generally less likely than women to self-identify as *white*. Yet, once we account for appearance, this pattern reverses: men become slightly more likely to report themselves as *white*. This is the only variable for which controlling for appearance reverses the direction of the correlation. It is important to note, however, that this reversal is not robust to the alternative face-detection algorithm discussed in Appendix C and should therefore be interpreted with caution. All other results reported in this paper remain robust to the alternative algorithm.

Marital status also shows a notable pattern: both married and single individuals are less likely to self-identify as *white* compared to formerly married individuals, though these effects weaken after accounting for appearance. Similarly, education is positively associated with identifying as *white*, but the magnitude of the effect is reduced once appearance is controlled for. Notably, even conditional on phenotypical appearance, college-educated individuals are still more likely to self-identify as *white* — specifically, 8.6 percentage points more likely than those with less than primary education. This finding supports the "social

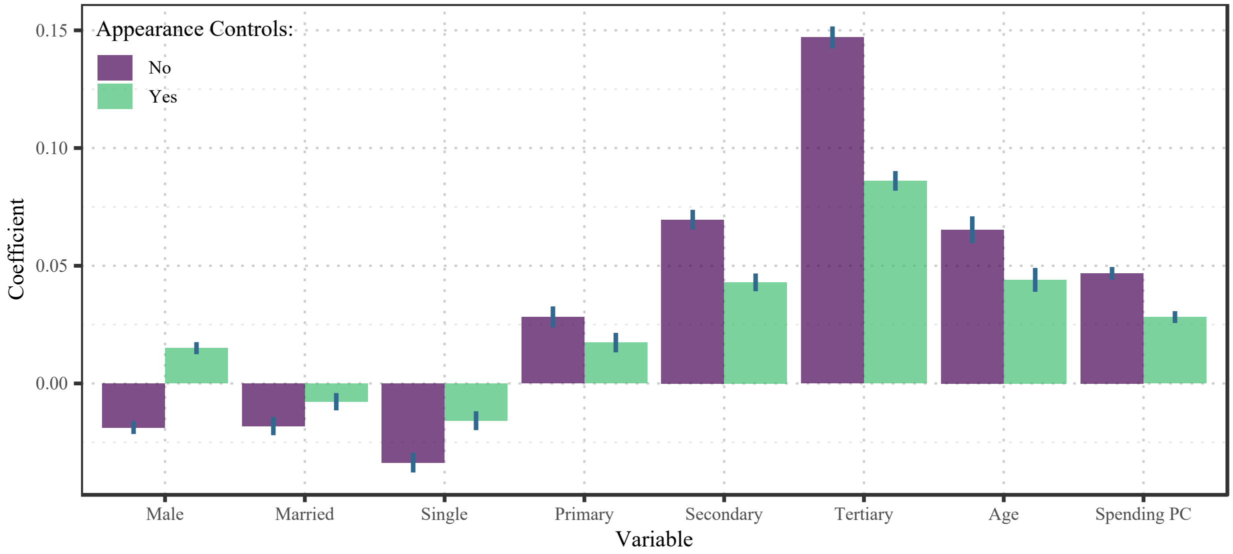

Figure 1: **Phenotypical Appearance Bias**. Marginal effects of the variables shown on the horizontal axis estimated by a logistic regression controlling for municipality fixed effects. The reference group for *Male* is *Female*; for *Marital Status*, it is *Separated/Divorced/Widowed*; and for *Education*, it is less than primary. Vertical line at the end of columns represent 95% confidence intervals.

whitening" hypothesis: upward socioeconomic mobility increases the likelihood of identifying as *white*, even when physical appearance remains constant. Lastly, older and wealthier individuals (proxied by campaign spending) also show a greater probability of reporting as *white*, although these effects also diminish when conditioning on appearance.

To assess whether the patterns described above are influenced by individuals whose appearance is harder to classify, we examine how the estimated relationships vary with the algorithm's confidence in identifying someone as *white*. The idea is that people with more ambiguous phenotypical traits may have greater room to shape how they report their racial identity. Figure 2 presents this conditional marginal effect distribution. Panels A, B, and C show how the influence of socioeconomic variables on self-reported race changes as the probability of being classified as *white* increases relative to the probability of being classified as *Latino/Hispanic*, holding all else constant. The results reveal that marginal effects decline significantly as individuals are more clearly perceived as *white* by the algorithm. In other words, socioeconomic characteristics matter more when an individual's appearance is harder to classify. When *Latino/Hispanic* is the reference group, the marginal effects peak around zero, consistent with the fact that this group is phenotypically diverse and often ambiguous. In contrast, panels D, E, and F use *black* as the reference group. Here, the peak occurs around 0.2, indicating that individuals with more mixed or intermediate appearances rely

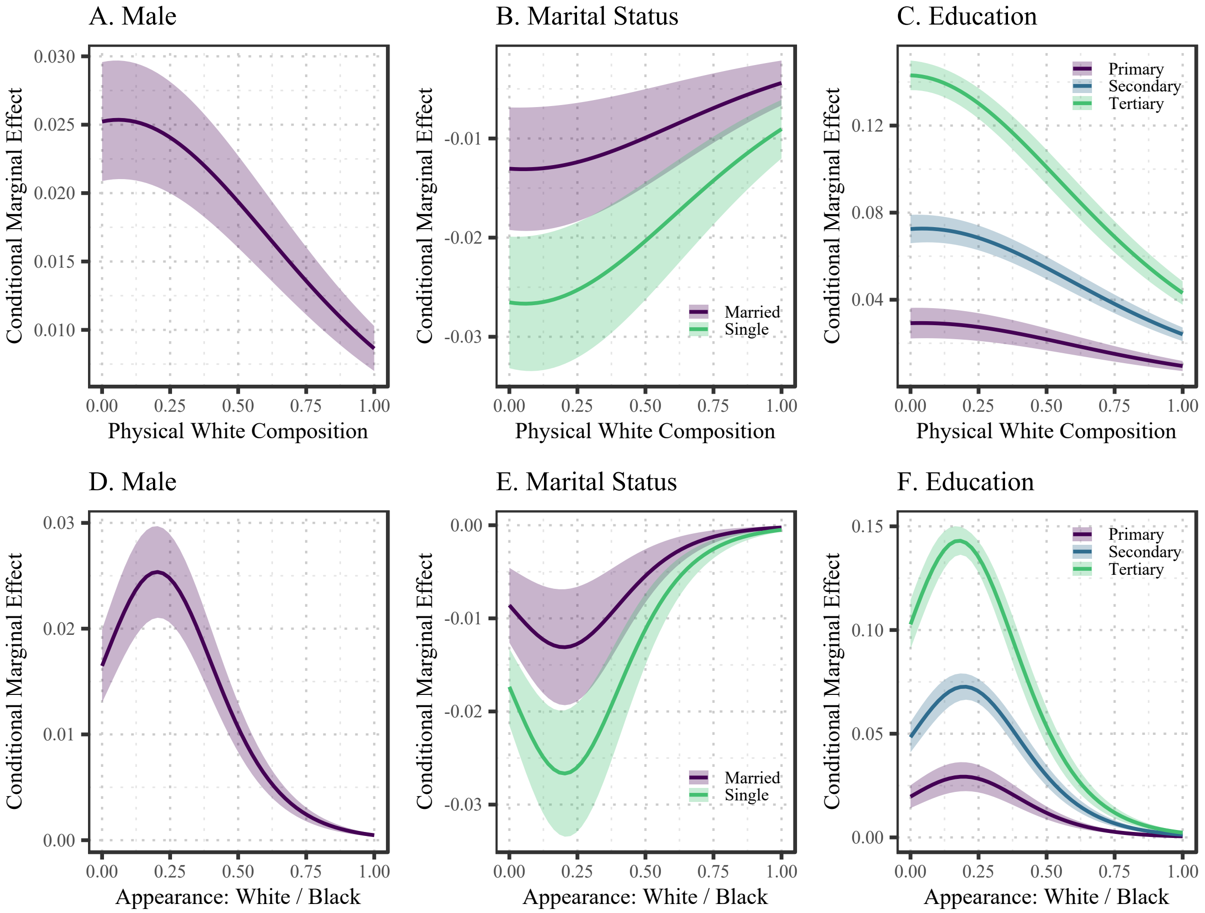

Figure 2: **Conditional Marginal Effects.** Marginal effect of the variables in the title, conditional on the probability of being classified as *white* by the algorithm. In Panels A to C, *Latino/Hispanic* is the reference group; in the remaining panels, the reference group is *black*. The reference group for *Male* is *Female*; for *Marital Status*, it is *Separated/Divorced/Widowed*; and for *Education*, it is less than primary.

more on socioeconomic cues when reporting race. As the probability of being classified as *black* increases, the influence of socioeconomic variables on racial self-identification diminishes.

Here, we have shown that individual characteristics are key to understanding how people identify racially. Phenotypical appearance, captured by the probabilities generated by our algorithm, strongly influences self-reported race, while socioeconomic conditions also help shape these identities. Moreover, the substantial increase in explanatory power once municipality fixed effects are included, reflected in the higher R^2 of our regressions, suggests that local context and its historical legacy play an important role in how racial boundaries are socially perceived. These patterns motivate our next step: to examine more closely the role of community-level perception in shaping racial identification.

5 Social Racial Perception Bias

To capture social perceptions about race, we extract municipality fixed effects from a logistic regression where the dependent variable is a binary indicator of whether an individual self-identifies as *white*. The explanatory variables include the probabilities assigned by the machine learning algorithm and the individual's socioeconomic characteristics described above. We interpret these fixed effects, expressed on the log-odds scale, as capturing residual municipal bias in racial classification—that is, the extent to which individuals with similar phenotypical appearance and socioeconomic profiles are more or less likely to self-identify as *white* depending on their municipality. We refer to this measure as the social racial perception bias, or simply social bias. In the analysis below, we use standardized values of this measure.

Panel A of Figure 3 displays the spatial distribution of social bias across Brazil. We observe clear differences in the municipality fixed effects between the South and Southeast regions compared to the rest of the country. These regions are more developed, featuring higher levels of education and income per capita. Notably, the South in particular experienced significant waves of European immigration in the late nineteenth century, driven by state policies aimed at promoting the "whitening" of Brazil's population (Dos Santos, 2002). It is thus plausible that the social perceptions formed during this historical period have persisted and continue to influence the social racial perception bias observed today.

Motivated by the observed relationship between historical formation and social bias, Panel B of Figure 3 illustrates the relationship between the social bias and the share of the slave population in 1872—the year of Brazil's first census, which recorded total and slave populations by municipality. To ensure consistent geographic boundaries over time, we aggregate the 1872 data to minimum comparable areas, assigning the same slave population share to current municipalities within each area. Our findings show that municipalities exhibiting a stronger social bias favoring *white* classification tend to have had larger share of slaves in the past. This association suggests that social bias is deeply rooted in cultural factors that also contribute to economic disparities along racial lines. In Table 2, we estimate a linear model using OLS to assess the relationship between the share of slave population and contemporary social bias.

Column (1) shows that a 10 percentage point increase in the share of the slave population in 1872 is associated with an increase of about 0.3 standard deviations in social bias. In column (2), we include state fixed effects to compare municipalities within the same state that likely faced similar policies over this century-long period. The coefficient remains very similar but is estimated more precisely. In column (3), we

Table 2: **Long-run Determinants of Social Racial Perception Bias**

	(1)	(2)	(3)	(4)
Dependent Variable: Municipality Fixed Effect				
Share of Slaves (1872)	2.95**	2.60***	2.12***	0.690***
	(1.24)	(0.880)	(0.736)	(0.267)
Share of White (2022)				3.33***
				(0.179)
Share of Male (2022)				0.693
				(0.488)
Share of Literate (2022)				0.391
				(0.306)
Wage Gap (2010)				-0.203
				(0.171)
State Fixed Effect		✓	✓	✓
Geography Controls			✓	✓
Observations	5,355	5,355	5,355	5,355
R^2	0.03124	0.58295	0.59800	0.70356

Spatially robust standard errors (Conley, 1999) with a 250km cutoff in parentheses. Social racial perception bias is the standardized value of municipalities fixed effects estimated from the logistic regression in column (3) of Table A.3. It measures the propensity of individuals in a municipality to self-report as white given their individual characteristics. Share of Slaves (1872) is the number of slaves over population as measured in the 1872 census. This variable is at the level of minimum comparable areas of 1872. Geography controls include the median temperature, median elevation, median rainfall, median terrain ruggedness index, distance to rivers (in log), distance to the coast (in log), and area (in log). Wage gap is the the nonwhite-to-white average wage ratio in 2010. * p < 0.10, ** p < 0.05, *** p < 0.01

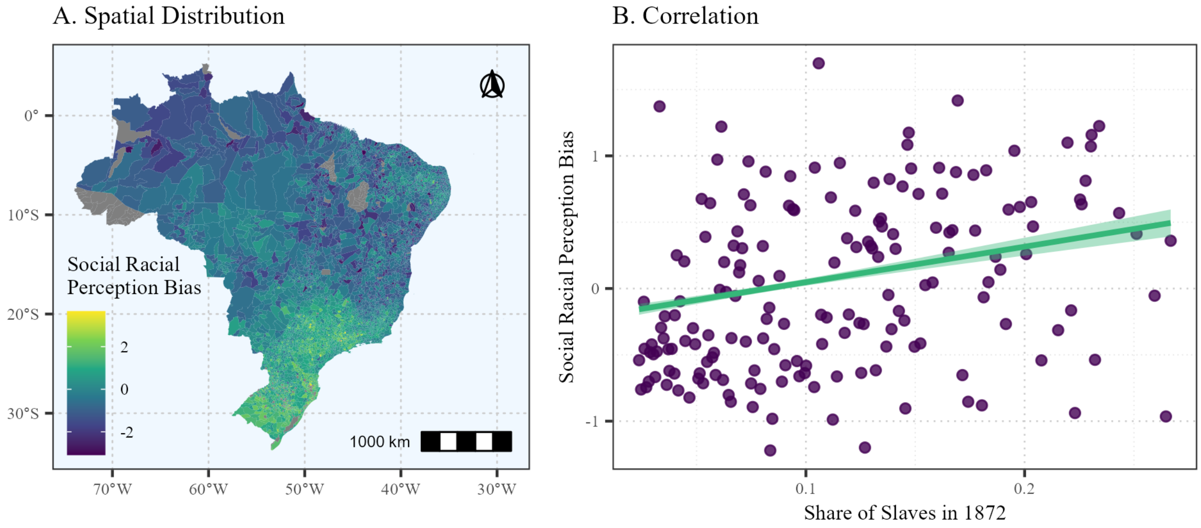

Figure 3: **Social racial perception bias.** Social racial perception bias are standardized municipality fixed effects estimated from the logit regression in column (3) of Table A.3. Panel A displays their spatial distribution, while Panel B shows their correlation with the share of slaves in 1872 in the municipality (associated with a minimum comparable are of 1872). Each point is a group of 300 municipalities with similar social bias with averaged share of slaves.

add a set of geographic control variables, including median temperature, median rainfall, median elevation, median Terrain Ruggedness Index, distance to rivers, distance to the coast, and area. These variables may correlate with the demand for slaves or other factors that shaped the development paths of municipalities and, consequently, their social perceptions of race. Nonetheless, the coefficient only decreases slightly and remains statistically significant. Overall, the results indicate a robust positive relationship between the share of slaves in 1872 and social racial perception bias today.

In column (4) of Table 2, we add the municipality-level controls: the share of *white* individuals, the share of males, the share of literate individuals from the 2022 Census, and the *nonwhite*-to-*white* wage gap in 2010. The concern is that, since our sample is composed of politicians, their behavior may closely follow the behavior of the electorate. For example, in areas where the population is (reportedly) more *white*, politicians may tend to self-classify as *white*. We refer to this as *herd behavior* (Banerjee, 1992).[4] Moreover, areas with a larger *nonwhite*-to-*white* wage gap may exhibit stronger social bias through aspirations for upward mobility, as discussed in the literature. Note, however, that rather than representing omitted variables that

[4] One concern with this measure is that the dependent variable may affect the share of reportedly *white* population. Although it is true that these variables capture the same response, the share of politicians in the population is negligible and unlikely to directly influence this control variable.

render our estimates spurious, these measures are possible channels through which colonial legacy manifests in today's biases in racial identity.

The results show that only the share of *white* individuals strongly predicts the social perception bias, suggesting that *herd behavior* is at play. Upward mobility aspirations, as measured by the *nonwhite*-to-*white* wage ratio, do not significantly explain the social perception bias. Importantly, the coefficient associated with the share of slaves in 1872 decreases significantly, indicating that *herd behavior* accounts for a substantial portion of the effect of slavery intensity on social perception bias. Nevertheless, this effect remains positive and statistically significant, implying that slavery intensity influences social perception bias through additional channels, which we discuss below.

Although the use of social racial perception bias at the municipality level is useful for observing general patterns and spatial distribution, we also examine the impact of the share of the slave population in 1872 on the probability of self-identifying as white, conditional on individual and municipal characteristics. We present the marginal effects at different levels of predicted probability of being white based on the photograph, with *black* as the reference category, in Figure 4. Detailed estimates of average effects are reported in Table A.4. Panel A shows the marginal effects controlling for phenotypical appearance and individual socioeconomic characteristics (gender, marital status, education, age, and campaign spending). Panel B additionally controls for geographical and socioeconomic characteristics of the municipality, listed in Table 2. All specifications include state fixed effects, and standard errors are clustered at the level of the 1872 minimum comparable areas.

In general, the results align with those discussed in Table 2. A larger share of slaves in 1872 is associated with a higher probability of self-identifying as *white*. Moreover, the effect diminishes when we control for variables that capture *herd behavior*. Importantly, the effect exhibits a hump-shaped relationship with the probability of being classified as *white* by the algorithm. Depending on the specification, the marginal effect peaks at around 15% or 25%, and in all cases tends to zero as the probability of being classified as *white* approaches one. This pattern suggests that variables influencing social racial perception act primarily on individuals with mixed phenotypical appearance, which is largely expected and reinforces that our method captures individuals' behavior regarding their racial identity.

Overall, the findings presented in this section provide strong evidence that historical factors——particularly the legacy of slavery——continue to shape contemporary social perceptions of race in Brazil. The persistent social bias in racial classification across municipalities appears to be deeply rooted in long-run demographic

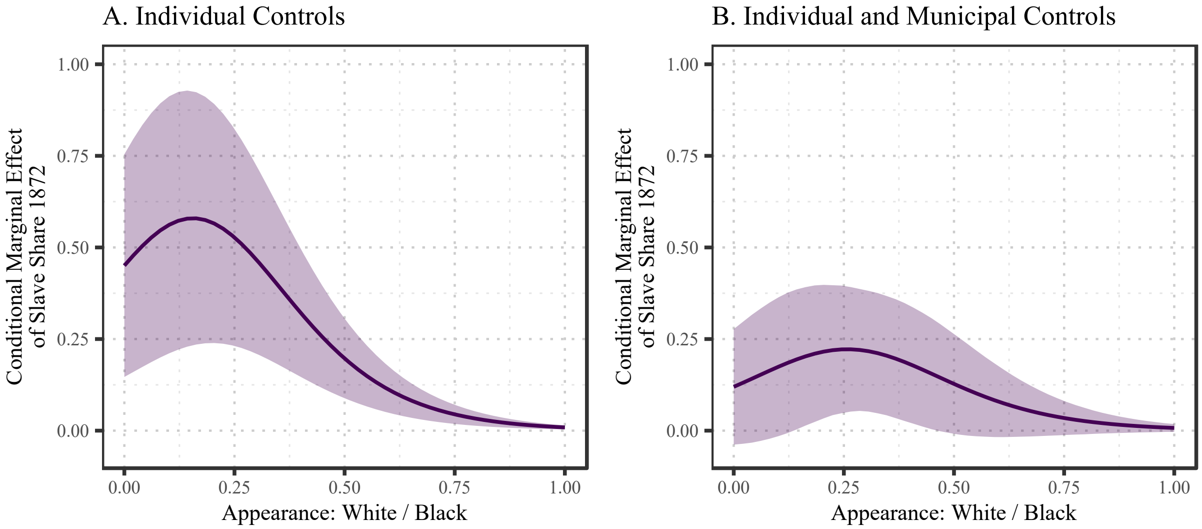

Figure 4: **Conditional Marginal Effects of Slave Share in 1872.** Marginal effect of the share of slaves in 1872, conditional on the probability of being classified as *white* by the algorithm. The reference group is *black*. Individual variables include phenotypical appearance probabilities, binary variables indicating if the individuals is male, married, single, and their education attainment, and continuous variables measure age and campaign spending. Municipality controls include the median temperature, median elevation, median rainfall, median terrain ruggedness index, distance to rivers (in log), distance to the coast (in log), area (in log), the share of white, the share of male and the share of literate individuals in the 2022 Census, and the nonwhite-to-white wage gap in 2010.

and institutional processes. Understanding this historical dimension is crucial for addressing current racial inequalities, as social perceptions both reflect and reinforce economic and social disparities.

The Role of Institutions The relationship between slavery intensity and social racial perception bias can be explained by several factors. For example, African slavery in Brazil fostered social norms that viewed *nonwhite* individuals as inferior to *white* individuals. It is plausible that these norms were stronger and more likely to persist over time in areas with a greater presence of slavery. Alternatively, a larger slave presence is also associated with the extraction of natural resources, which often led to the introduction of extractive institutions—institutions that can persist over time (Acemoglu, Johnson and Robinson, 2002). In contrast, areas of European settlement are associated with more inclusive institutions.

To investigate the channels proposed above, we compare the effects of slavery intensity on racial self-classification with the effects of state-sponsored European settlements in the state of São Paulo. To this end, we closely follow Rocha et al. (2017), who find that these settlements are associated with persistent investments in education, leading to long-lasting benefits in terms of human capital. European settlements were sponsored by the government to replace the declining slave labor force with “higher quality” workers from Europe.[5] Many European migrants directly replaced slave workers on already established coffee plantations, while some immigrants were granted land in previously unused areas.[6]

If social norms depicting African slaves as inferior are driving our previous results, we should expect European settlements to have no effect, or similar effects, on racial identity today, since these colonies were founded on the assumption that European workers were superior. On the other hand, if the result is driven by the introduction of extractive *versus* inclusive institutions, we should expect state-sponsored settlements to have the opposite effect on racial identity compared to slavery intensity. The results are presented in Figure 5.

Figure 5 shows the marginal effects of the share of slaves in 1872 and the presence of European settlements in 1920, conditional on the probability of being classified as *white* by the algorithm, with *black* as the reference group. The sample includes only individuals from the state of São Paulo. Panels A and C control for individual characteristics as in Figure 4, while Panels B and D add municipality characteristics as in Figure 5. Following Rocha et al. (2017), we include the following predetermined controls from the 1872 Census among the municipality characteristics: share of children in school, share of foreign residents, share of literate individuals, share of agricultural workers, share of workers in industry, public servants per capita, teachers per capita, legal professionals per capita, and the log of population density. Detailed average effects are presented in Table A.5.

The results show that the effect of slavery intensity in 1872 on racial self-classification in this restricted sample is similar to the effect in the full sample depicted in Figure 4. Importantly, the effects of state-sponsored European settlements are the opposite: individuals—particularly those with mixed appearance—are more likely to self-identify as *nonwhite* if they live in municipalities where state-sponsored European settlements existed in 1920. It is important to note that these estimates control for the share of slaves in 1872, so the effect is not simply driven by a lower share of slaves.

These results lead to the conclusion that the historical effect of slavery on social racial perception bias is driven by the introduction of extractive institutions, in contrast to inclusive institutions introduced by European settlements.

[5] The main reason for the continuous decline in the slave labor force was the passage of a national law forbidding the international trade of slaves in 1850.

[6] Although land parcels were assigned to European immigrants and initial costs covered by the state, contracts required reimbursement of the state in a series of payments starting after the first harvest. See Rocha et al. (2017) for details.

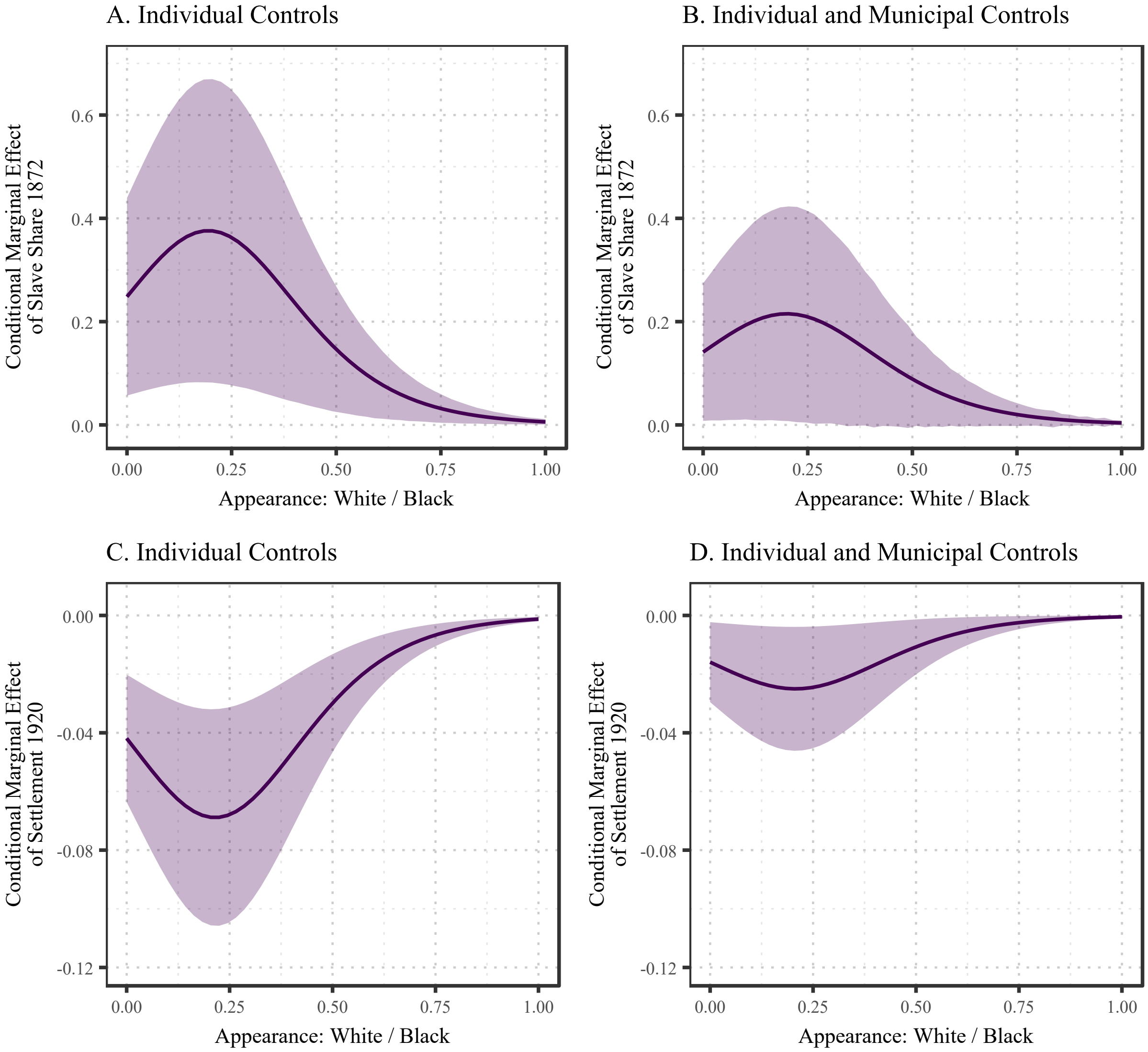

Figure 5: **Conditional Marginal Effects of Slave Share in 1872 and European Settlements in the State of São Paulo.** Marginal effect of the share of slaves in 1872 and the existence of a state-sponsored European settlement in 1920, conditional on the probability of being classified as *white* by the algorithm. The reference group is *black*. Individual variables include phenotypical appearance probabilities, binary variables indicating if the individuals is male, married, single, and their education attainment, and continuous variables measure age and campaign spending. Municipality controls include the median temperature, median elevation, median rainfall, median terrain ruggedness index, distance to rivers (in log), distance to the coast (in log), area (in log), the share of white, the share of male and the share of literate individuals in the 2022 Census, the nonwhite-to-white wage gap in 2010, and the following 1872 variables: share of children in school, share of foreign residents, share of literate individuals, share of agricultural workers, share of workers in industry, public servants per capita, teachers per capita, legal professionals per capita, and the log of population density.

6 Conclusion

This paper provides new evidence on the complex relationship between phenotypical appearance, socioeconomic characteristics, and social context in shaping racial identity in Brazil. Using a novel approach that leverages machine learning-based phenotypical classification alongside self-reported race from a large sample of political candidates, we show that while physical appearance strongly predicts racial self-identification, socioeconomic factors and community-level social perceptions also play significant and independent roles. In particular, our results support the "social whitening" hypothesis, whereby upward socioeconomic mobility increases the likelihood of identifying as *white*, even conditional on appearance. Additionally, social bias at the municipal level—captured through fixed effects—accounts for a substantial share of variation in racial self-identification, highlighting the importance of local cultural norms and perceptions in the construction of race.

We further uncover that these social biases have deep historical roots linked to the legacy of slavery, with municipalities that had higher shares of slave populations in 1872 displaying stronger social bias toward whiteness today. This historical persistence underscores the importance of considering long-term institutional and cultural factors when studying racial identity and inequality. Together, our findings contribute to a more nuanced understanding of race as a social construct shaped by appearance, economic status, and collective historical experience. Future research and policy efforts aiming to address racial inequality in Brazil should take into account these multifaceted determinants, recognizing that racial identity is embedded within both individual traits and broader societal dynamics.

Appendix A Omitted Tables

Table A.1: **Descriptive Statistics**

Variables	Count	Mean	Std. Dev.	Min	Max
Physical Appearance (from photo)					
White	535,832	0.2847	0.2767	0	1
Asian	535,832	0.0978	0.1658	0	1
Indian	535,832	0.0681	0.0788	0	1
Black	535,832	0.0636	0.1801	0	1
Middle Eastern	535,832	0.1526	0.1747	0	1
Latino/Hispanic	535,832	0.3333	0.2135	0	1
Sociodemographic (reported)					
White	558,226	0.4808	0.4996	0	1
Male	558,226	0.6646	0.4721	0	1
Single	558,226	0.373	0.4836	0	1
Age	558,220	45.587	11.5599	16	99
Highest Education Attained					
Primary	558,226	0.1717	0.3771	0	1
Secondary	558,226	0.4257	0.4945	0	1
Terciary	558,226	0.2433	0.4291	0	1

Table A.2: **Race classification by *DeepFace***

Individual	Asian	Indian	Black	White	Middle Eastern	Latino Hispanic
Angela Merkel (German Politician)	0.00	0.00	0.00	99.97	0.01	0.02
Aung San Suu Kyi (Burmese Politician)	7.22	12.62	1.33	22.97	27.57	28.30
Benjamin Netanyahu (Israeli Politician)	0.39	5.53	0.15	20.53	59.14	14.27
Bill Clinton (American Policitian)	0.14	0.06	0.00	88.38	5.42	5.99
Bill Gates (American Businessperson)	0.00	0.00	0.00	98.71	0.83	0.45
Condoleezza Rice (American Policitian)	61.42	4.91	12.18	0.97	0.39	20.14
Donald Trump (American Policitian)	0.00	0.00	0.00	99.98	0.01	0.01
Elizabeth Warren (American Policitian)	3.14	0.09	0.03	63.10	2.41	31.22
George W Bush (American Policitian)	0.86	1.02	0.12	64.71	8.03	25.26
George Clooney (American Actor)	0.03	0.01	0.00	95.96	1.89	2.12
Hillary Clinton (American Policitian)	0.01	0.00	0.00	95.79	1.02	3.18
Hu Jintao (Chinese Politician)	93.37	0.22	0.03	3.88	0.71	1.80
James Dimon (American Businessperson)	3.91	0.30	0.06	66.14	5.76	23.83
Jeff Bezos (American Businessperson)	87.89	1.83	0.09	3.03	0.15	7.01
Joe Biden (American Policitian)	0.40	0.03	0.00	96.24	1.56	1.77
Kim Jong Un (North Korean Politician)	99.95	0.01	0.00	0.00	0.00	0.04
Christine Lagarde (French Politician)	0.03	0.01	0.00	90.27	1.41	8.28
LeBron James (American Basketball Player)	1.20	0.06	98.64	0.05	0.01	0.04
Mark Zuckenberg (American Businessperson)	0.00	0.00	0.00	100.00	0.00	0.00
Michelle Obama (American Author)	7.49	6.08	11.56	48.62	10.24	16.00
Nancy Pelosi (American Policitian)	0.00	0.00	0.00	99.34	0.41	0.25
Oprah Winfrey (American Talk Show Host)	0.09	2.04	97.02	0.00	0.00	0.84
Pope Francis	0.06	0.35	0.02	74.49	15.42	9.67
Barack Obama (American Policitian)	3.08	7.63	61.83	3.29	2.78	21.40
Tayyip Erdogan (Turkish Politician)	22.38	16.25	7.68	12.96	13.67	27.07
Janet Yellen (American Economist)	0.00	0.00	0.00	99.99	0.00	0.01
Narendra Modi (Indian Politician)	51.34	9.22	2.24	8.56	23.91	4.74
Elon Musk (American Businessperson)	12.07	0.09	0.01	80.28	1.17	6.37
Steve Jobs (American Businessperson)	0.03	0.08	0.00	88.18	4.81	6.89
Tim Cook (American Businessperson)	0.00	0.00	0.00	100.00	0.00	0.00
Vladimir Putin (Russian Politician)	0.12	0.06	0.01	97.78	1.36	0.67
Xi Jinping (Chinese Politician)	96.16	0.86	0.00	0.03	0.00	2.95

Table A.3: **Racial Identify and Socioeconomic Characteristics**

	OLS		Logit	
	(1)	(2)	(3)	(4)
Dependent Variable: Reported White				
Male	0.024***	0.014***	0.101***	0.015***
	(0.001)	(0.001)	(0.009)	(0.001)
Married	-0.025***	-0.009***	-0.052***	-0.008***
	(0.002)	(0.002)	(0.013)	(0.002)
Single	-0.067***	-0.017***	-0.107***	-0.016***
	(0.003)	(0.002)	(0.014)	(0.002)
Primary Educ.	0.033***	0.016***	0.117***	0.017***
	(0.003)	(0.002)	(0.014)	(0.002)
Secondary Educ.	0.047***	0.042***	0.291***	0.043***
	(0.003)	(0.002)	(0.013)	(0.002)
Tertiary Educ.	0.092***	0.086***	0.576***	0.086***
	(0.004)	(0.002)	(0.014)	(0.002)
Age (log)	0.074***	0.043***	0.297***	0.044***
	(0.003)	(0.003)	(0.017)	(0.003)
Spending pc (log)	0.006	0.029***	0.190***	0.028***
	(0.003)	(0.001)	(0.009)	(0.001)
Appearance Controls	✓	✓	✓	✓
Municipality Fixed Effect		✓	✓	✓
Average Marginal Effect	✓	✓		✓
Observations	516,189	516,189	511,446	511,446
(Pseudo) R^2	0.22911	0.39503	0.34480	0.34480

Standard errors clustered at the municipality level are reported in parentheses. The dependent variable is a binary indicator equal to 1 if the individual self-reported as white. Male is a binary indicator equal to 1 if the individual is male. Married and Single indicate marital status, with divorced/separated/widowed as the reference group. Primary, Secondary, and Tertiary Educ. denote the highest level of education attained, with less than primary education as the reference group. Spending pc refers to the candidate's campaign spending divided by the municipality's population. Appearance controls are the algorithm-assigned probabilities that the individual is identified as each respective race. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.4: **Slavery and Social Racial Perception Bias - Individual Level Regressions**

	(1)	(2)	(3)	(4)
Dependent Variable: Reported White				
Share of Slaves (1872)	0.342**	0.376***	0.375***	0.139**
	(0.138)	(0.114)	(0.120)	(0.056)
Individual Controls		✓	✓	✓
Geography Controls			✓	✓
Municipality Controls				✓
Observations	551,732	510,200	510,200	510,200
Pseudo R^2	0.15497	0.29252	0.29472	0.31308

Standard errors clustered at the level of minimum comparable areas of 1872. The dependent variable is a binary indicator equal to 1 if the individual self-reported as white. Share of Slaves (1872) is the number of slaves over population as measured in the 1872 census. This variable is at the level of minimum comparable areas of 1872. Individual variables include phenotypical appearance probabilities, binary variables indicating if the individuals is male, married, single, and their education attainment, and continuous variables measure age and campaign spending. Geography controls include the median temperature, median elevation, median rainfall, median terrain ruggedness index, distance to rivers (in log), distance to the coast (in log), and area (in log). Municipality controls include the share of white, the share of male and the share of literate individuals in the 2022 Census, and the nonwhite-to-white wage gap in 2010. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: **European Settlements and Social Racial Perception Bias - Individual Level Regressions**

	(1)	(2)	(3)	(4)
Dependent Variable: Reported White				
European Settlements	-0.047***	-0.033***	-0.023**	-0.015**
	(0.017)	(0.012)	(0.011)	(0.007)
Share of Slaves (1872)			0.308***	0.130**
			(0.117)	(0.064)
Individual Controls		✓	✓	✓
Geography Controls			✓	✓
Municipality Controls				✓
Observations	93,852	87,561	85,930	85,930
Pseudo R^2	0.00164	0.22902	0.23797	0.24840

Standard errors clustered at the level of minimum comparable areas of 1872. The dependent variable is a binary indicator equal to 1 if the individual self-reported as white. Share of Slaves (1872) is the number of slaves over population as measured in the 1872 census. This variable is at the level of minimum comparable areas of 1872. Individual variables include phenotypical appearance probabilities, binary variables indicating if the individuals is male, married, single, and their education attainment, and continuous variables measure age and campaign spending. Geography controls include the median temperature, median elevation, median rainfall, median terrain ruggedness index, distance to rivers (in log), distance to the coast (in log), and area (in log). Municipality controls include the share of white, the share of male and the share of literate individuals in the 2022 Census, the nonwhite-to-white wage gap in 2010 and the following 1872 variables: share of children in school, share of foreign residents, share of literate individuals, share of agricultural workers, share of workers in industry, public servants per capita, teachers per capita, legal professionals per capita, and the log of population density. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix B Wealth Sample

Candidates in Brazilian elections are required to disclose the value of their assets, which we use as a proxy for wealth. This measure has important limitations. There is no formal verification of the reported values, although physical assets such as houses and vehicles may be visible and subject to public scrutiny. In contrast, bank accounts are not publicly observable and are more likely to be omitted. As a result, many candidates report zero wealth. Souto-Maior and Borba (2019) suggests that some candidates may believe only high-value assets must be declared or that those exempt from income taxes are not required to report, leading to underreporting.

We replicate the results in Tables 2, A.4 and A.5 using only individuals who reported positive wealth. As shown in Table B.1, the positive relationship between wealth and the probability of self-identifying as *white* remains robust. The coefficients for other covariates change little. In Table B.2, columns 1–3 reproduce the estimates from Table A.4 and columns 4–6 from Table A.5. Using this restricted sample and controlling for wealth slightly strengthens the effects of slavery and European settlement on the probability of self-identifying as *white*. Overall, our conclusions remain unchanged when wealth is included as a control.

Table B.1: **Racial Identify and Socioeconomic Characteristics - Wealth Control**

	OLS		Logit	
	(1)	(2)	(3)	(4)
Dependent Variable: Reported White				
Male	0.025***	0.014***	0.113***	0.016***
	(0.002)	(0.002)	(0.012)	(0.002)
Married	-0.024***	-0.010***	-0.065***	-0.009***
	(0.003)	(0.002)	(0.016)	(0.002)
Single	-0.058***	-0.015***	-0.098***	-0.014***
	(0.003)	(0.003)	(0.018)	(0.003)
Primary Educ.	0.026***	0.013***	0.093***	0.013***
	(0.003)	(0.003)	(0.019)	(0.003)
Secondary Educ.	0.038***	0.036***	0.261***	0.037***
	(0.004)	(0.002)	(0.017)	(0.002)
Terciary Educ.	0.080***	0.081***	0.556***	0.081***
	(0.004)	(0.003)	(0.018)	(0.003)
Age (log)	0.066***	0.038***	0.266***	0.038***
	(0.005)	(0.003)	(0.024)	(0.003)
Wealth (log)	0.013***	0.013***	0.087***	0.013***
	(0.0007)	(0.0005)	(0.003)	(0.0005)
Appearance Controls	✓	✓	✓	✓
Municipality Fixed Effect		✓	✓	✓
Average Marginal Effect	✓	✓		✓
Observations	327,770	327,770	321,617	321,617
(Pseudo) R^2	0.23268	0.41077	0.35697	0.35697

Standard errors clustered at the municipality level are reported in parentheses. The dependent variable is a binary indicator equal to 1 if the individual self-reported as white. Male is a binary indicator equal to 1 if the individual is male. Married and Single indicate marital status, with divorced/separated/widowed as the reference group. Primary, Secondary, and Tertiary Educ. denote the highest level of education attained, with less than primary education as the reference group. Wealth is the value of individual assets reported by candidates. Appearance controls are the algorithm-assigned probabilities that the individual is identified as each respective race. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.2: **Long-run Determinants of Social Racial Perception Bias - Wealth Control**

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable: Reported White						
Share of Slaves (1872)	0.427***	0.399***	0.130**	0.190**	0.287**	0.118*
	(0.133)	(0.130)	(0.065)	(0.085)	(0.114)	(0.067)
European Settlements				-0.044***	-0.026*	-0.016***
				(0.012)	(0.014)	(0.006)
Individual Controls	✓	✓	✓	✓	✓	✓
Geography Controls		✓	✓		✓	✓
Municipality Controls			✓			✓
Observations	324,034	324,034	324,034	54,369	53,383	53,383
Pseudo R^2	0.29892	0.30135	0.32189	0.23106	0.23930	0.25201

Standard errors clustered at the level of minimum comparable areas of 1872. The dependent variable is a binary indicator equal to 1 if the individual self-reported as white. Share of Slaves (1872) is the number of slaves over population as measured in the 1872 census. This variable is at the level of minimum comparable areas of 1872. Individual variables include phenotypical appearance probabilities, binary variables indicating if the individuals is male, married, single, and their education attainment, and continuous variables measure age and wealth. Geography controls include the median temperature, median elevation, median rainfall, median terrain ruggedness index, distance to rivers (in log), distance to the coast (in log), and area (in log). Municipality controls include the share of white, the share of male and the share of literate individuals in the 2022 Census, the nonwhite-to-white wage gap in 2010 and the following 1872 variables: share of children in school, share of foreign residents, share of literate individuals, share of agricultural workers, share of workers in industry, public servants per capita, teachers per capita, legal professionals per capita, and the log of population density. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix C Retina Face Algorithm

In this section, we present results obtained with an alternative face-detection algorithm, *RetinaFace* (Deng, Guo, Zhou, Yu, Kotsia and Zafeiriou, 2019; Serengil and Ozpinar, 2024). Using this model increases the sample size by about 14,000 observations. The coefficients for socioeconomic characteristics reported in Table C.1 are nearly identical to those estimated with the *opencv* algorithm in Table A.3, with one notable exception: the coefficient on sex, which changes sign from positive to negative. This indicates that the sex result is not robust and should be interpreted with caution. The findings on the long-run determinants of racial identity, shown in Table C.2, are likewise very similar when using *RetinaFace*.

Table C.1: **Racial Identify and Socioeconomic Characteristics - Retina Face Algorithm**

	OLS		Logit	
	(1)	(2)	(3)	(4)
Dependent Variable: Reported White				
Male	-0.017***	-0.013***	-0.052***	-0.008***
	(0.001)	(0.001)	(0.009)	(0.001)
Married	-0.025***	-0.009***	-0.054***	-0.008***
	(0.002)	(0.002)	(0.012)	(0.002)
Single	-0.066***	-0.017***	-0.105***	-0.015***
	(0.003)	(0.002)	(0.013)	(0.002)
Primary Educ.	0.032***	0.016***	0.118***	0.017***
	(0.003)	(0.002)	(0.014)	(0.002)
Secondary Educ.	0.045***	0.041***	0.288***	0.042***
	(0.003)	(0.002)	(0.013)	(0.002)
Terciary Educ.	0.090***	0.086***	0.570***	0.085***
	(0.004)	(0.002)	(0.014)	(0.002)
Age (log)	0.082***	0.050***	0.360***	0.053***
	(0.003)	(0.003)	(0.017)	(0.003)
Spending pc (log)	0.006*	0.028***	0.185***	0.027***
	(0.003)	(0.001)	(0.009)	(0.001)
Appearance Controls	✓	✓	✓	✓
Municipality Fixed Effect		✓	✓	✓
Average Marginal Effect	✓	✓		✓
Observations	530,997	530,997	526,347	526,347
(Pseudo) R^2	0.23895	0.39782	0.34853	0.34853

Standard errors clustered at the municipality level are reported in parentheses. The dependent variable is a binary indicator equal to 1 if the individual self-reported as white. Male is a binary indicator equal to 1 if the individual is male. Married and Single indicate marital status, with divorced/separated/widowed as the reference group. Primary, Secondary, and Tertiary Educ. denote the highest level of education attained, with less than primary education as the reference group. Spending pc refers to the candidate's campaign spending divided by the municipality's population. Appearance controls are the algorithm-assigned probabilities that the individual is identified as each respective race. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C.2: **Long-run Determinants of Social Racial Perception Bias - Retina Face Algorithm**

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable: Reported White						
Share of Slaves (1872)	0.353***	0.358***	0.127**	0.220**	0.304**	0.129**
	(0.113)	(0.120)	(0.057)	(0.096)	(0.120)	(0.065)
European Settlements				-0.045***	-0.025**	-0.017***
				(0.012)	(0.011)	(0.007)
Individual Controls	✓	✓	✓	✓	✓	✓
Geography Controls		✓	✓		✓	✓
Municipality Controls			✓			✓
Observations	524,898	524,898	524,898	89,982	88,304	88,304
Pseudo R^2	0.29671	0.29887	0.31703	0.24041	0.24816	0.25862

Standard errors clustered at the level of minimum comparable areas of 1872. The dependent variable is a binary indicator equal to 1 if the individual self-reported as white. Share of Slaves (1872) is the number of slaves over population as measured in the 1872 census. This variable is at the level of minimum comparable areas of 1872. Individual variables include phenotypical appearance probabilities, binary variables indicating if the individuals is male, married, single, and their education attainment, and continuous variables measure age and wealth. Geography controls include the median temperature, median elevation, median rainfall, median terrain ruggedness index, distance to rivers (in log), distance to the coast (in log), and area (in log). Municipality controls include the share of white, the share of male and the share of literate individuals in the 2022 Census, the nonwhite-to-white wage gap in 2010 and the following 1872 variables: share of children in school, share of foreign residents, share of literate individuals, share of agricultural workers, share of workers in industry, public servants per capita, teachers per capita, legal professionals per capita, and the log of population density. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

References

Acemoglu, Daron, Simon Johnson, and James A Robinson, "Reversal of fortune: Geography and institutions in the making of the modern world income distribution," *The Quarterly journal of economics*, 2002, *117* (4), 1231–1294.

Agadjanian, Alexander and Dean Lacy, "Changing votes, changing identities? Racial fluidity and vote switching in the 2012–2016 US Presidential Elections," *Public Opinion Quarterly*, 2021, *85* (3), 737–752.

Andrews, George Reid, "Brazilian Racial Democracy, 1900-90: An American Counterpoint," *Journal of Contemporary History*, 1996, *31* (3), 483–507.

Antman, Francisca M and Brian Duncan, "American Indian Casinos and Native American Self-Identification," *Journal of the European Economic Association*, 2023, *21* (6), 2547–2585.

_ and _ , "Ethnic Identity and Anti-Immigrant Sentiment: Evidence from Proposition 187," Technical Report, National Bureau of Economic Research 2024.

Bailey, Stanley R and Edward E Telles, "Multiracial versus collective black categories: examining census classification debates in Brazil," *Ethnicities*, 2006, *6* (1), 74–101.

Banerjee, Abhijit V, "A simple model of herd behavior," *The quarterly journal of economics*, 1992, *107* (3), 797–817.

Conley, Timothy G, "GMM estimation with cross sectional dependence," *Journal of econometrics*, 1999, *92* (1), 1–45.

de Lucena Coelho, Thiago, Fernanda Estevan, Marcos Nakaguma, and Alexandre Rabelo, "Do Black Politicians Matter? Political Leadership and Racial Composition in Top Public Sector Positions," 2024. Available at SSRN: https://ssrn.com/abstract=5183898.

Deng, Jiankang, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou, "RetinaFace: Single-stage Dense Face Localisation in the Wild," 2019. Available at https://arxiv.org/abs/1905.00641.

Derenoncourt, Ellora, Chi Hyun Kim, Moritz Kuhn, and Moritz Schularick, "Wealth of two nations: The US racial wealth gap, 1860–2020," *The Quarterly Journal of Economics*, 2024, *139* (2), 693–750.

Dupree-Wilson, Teisha, "Phenotypic proximity: Colorism and intraracial discrimination among Blacks in the United States and Brazil, 1928 to 1988," *Journal of Black Studies*, 2021, *52* (5), 528–546.

Francis, Andrew M and Maria Tannuri-Pianto, "Endogenous race in Brazil: affirmative action and the construction of racial identity among young adults," *Economic Development and cultural change*, 2013, *61* (4), 731–753.

Francis-Tan, Andrew and Maria Tannuri-Pianto, "Affirmative action in Brazil: global lessons on racial justice and the fight to reduce social inequality," *Oxford Review of Economic Policy*, 2024, *40* (3), 642–655.

Htun, Mala, "From "racial democracy" to affirmative action: changing state policy on race in Brazil," *Latin American Research Review*, 2004, *39* (1), 60–89.

IBGE, "Sintese de Indicadores Sociais: Uma análise da condições de vida da população brasileira," Technical Report, Instituto Brasileiro de Geografia e Estatística - IBGE 2024.

Karkkainen, Kimmo and Jungseock Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in "Proceedings of the IEEE/CVF winter conference on applications of computer vision" 2021, pp. 1548–1558.

Lang, Kevin and Ariella Kahn-Lang Spitzer, "Race discrimination: An economic perspective," *Journal of Economic Perspectives*, 2020, *34* (2), 68–89.

Laudares, Humberto and Felipe Valencia Caicedo, "Tordesillas, Slavery and the Origins of Brazilian Inequality," Technical Report, CEPR Discussion Papers 2023.

Mitchell, Gladys, "The politics of skin color in Brazil," *The Review of Black Political Economy*, 2010, *37* (1), 25–41.

Morning, Ann, *The Nature of Race: How Scientists Think and Teach about Human Difference*, Berkeley: University of California Press, 2011.

Naritomi, Joana, Rodrigo R Soares, and Juliano J Assunção, "Institutional development and colonial heritage within Brazil," *The journal of economic history*, 2012, *72* (2), 393–422.

Nunn, Nathan and Leonard Wantchekon, "The slave trade and the origins of mistrust in Africa," *American economic review*, 2011, *101* (7), 3221–3252.

Omi, Michael and Howard Winant, *Racial Formation in the United States*, 3rd ed., New York: Routledge, 2014.

Rocha, Rudi, Claudio Ferraz, and Rodrigo R Soares, "Human capital persistence and development," *American Economic Journal: Applied Economics*, 2017, *9* (4), 105–136.

Santos, Sales Augusto Dos, "Historical roots of the "Whitening" of Brazil," *Latin American Perspectives*, 2002, *29* (1), 61–82.

Saperstein, Aliya, "Recognizing Identity Fluidity in Demographic Research," *Population and Development Review*, 2025, *51* (1), 519–538.

Schwartzman, Luisa Farah, "Does Money Whiten? Intergenerational Changes in Racial Classification in Brazil," *American Sociological Review*, 2007, *72* (6), 940–963.

Serengil, Sefik and Alper Ozpinar, "A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules," *Journal of Information Technologies*, 2024, *17* (2), 95–107.

Souto-Maior, Cesar Duarte and José Alonso Borba, "Consistência na declaração de bens dos candidatos nas eleições brasileiras: ficção ou realidade?," *Revista de Administração Pública*, 2019, *53*, 195–213.

Telles, Edward and Tianna Paschel, "Who is black, white, or mixed race? How skin color, status, and nation shape racial classification in Latin America," *American Journal of Sociology*, 2014, *120* (3), 864–907.

Telles, Edward E, *Race in another America: The significance of skin color in Brazil*, Princeton University Press, 2004.