Training Data Diversity Shapes Racial Bias in AI Face Classification

Jeffrey J. Berg[1*], Anke Stoll[2*], Qianqian Cui[3], & David M. Amodio[2]

[1]New York University, [2]University of Amsterdam, and [3]University of California, Santa Barbara

*Shared first author

Corresponding author: David Amodio (d.m.amodio@uva.nl)

Keywords: Face, Algorithmic, Bias, AI, Race, Gender

Acknowledgement: This article is based in part on research conducted for Jeffrey Berg's doctoral dissertation at New York University.

Abstract

Can a lack of racial diversity in training data produce human-like forms of AI face classification bias? We examined the causal effect of racially-imbalanced training data on face classification model disparities and whether these disparities resemble three well-documented human biases—the *cross-race effect, hypodescent*, and the *gendered-race effect.* We further examined whether balanced training data mitigate these disparities. In three experiments, we trained deep Vision Transformer models for face classification on 16,000 face images, experimentally manipulating racial composition while holding other dataset features and training procedures constant. Study 1 showed that models trained on low-diversity (majority-White) datasets classified White faces more accurately than Black, East Asian, and Indian faces, resembling the human cross-race effect. In Study 2, using parametrically-morphed Black-White biracial face images to estimate race-classification thresholds, lower-diversity models required more Black than White content to reach this threshold, yet represented biracial faces with more Black content. Lastly, Study 3 revealed a gendered-race effect: regardless of training set imbalance, models classified Black men more accurately than Black women, but East Asian women more accurately than East Asian men. The use of racially-balanced training data nearly eliminated disparities in monoracial classification, reduced bias in biracial classification, but it did not alter the gendered-race effect. These findings show that training data composition causally shapes algorithmic performance in ways that both mimic and diverge from human biases. We discuss implications for AI-based discrimination and both the promise and limitations of data balancing as a strategy to mitigate algorithmic bias.

(max 250 words)

Significance Statement (50-120 words)

Do racially imbalanced training datasets produce systematic disparities in AI face classification? By experimentally manipulating training set racial composition while holding dataset size, image source, gender balance, model architecture, and evaluation procedures constant, we show that racial imbalance causally shapes model performance in ways that both resemble and diverge from well-documented human social-categorization biases. Balanced training data substantially reduced cross-race disparities in race classification accuracy but did not eliminate racial disparities in biracial classification thresholds or gender classification. These findings demonstrate the causal effects of training data diversity. Moreover, by revealing where AI classification patterns align with or diverge from human race perception, we identify potential risks for the acceptance and perpetuation of discriminatory AI outputs by human users.

Artificial intelligence (AI) systems are increasingly used to infer socially meaningful characteristics from human faces, such as race and gender, with applications ranging from identity verification and surveillance to policing, security, and marketing (1–3). Yet face classification systems have been found to exhibit systematic racial disparities, including lower accuracy for women and people with darker skin tones (4, 5). Such disparities can be especially harmful when AI outputs inform high-stakes decisions, such as in policing (6–8), and particularly when they align with existing patterns of human racial bias (9, 10). To date, however, the causal impact of training data imbalance on racial disparities in face classification, and their resemblance to human social perception, has not been investigated. Here, we experimentally tested whether racial imbalances in training data produce racial disparities in face classification model performance and, if so, whether these disparities mimic well-known human biases in face categorization.

Patterns of racial bias in AI face categorization systems

Demographic characteristics of faces—including race and gender—are known to influence the performance of AI systems for face detection, recognition, and classification, producing systematic disparities in model output (11–15). These disparities have been attributed, in part, to racial imbalances in the datasets used to train classification models (4, 5, 16–19). For example, many existing benchmark datasets (e.g., LFW+, IJB-A, Adience, CelebA) have been reported to contain high proportions of White faces (e.g., 80% or more) and uneven representations of other racial/ethnic groups, in addition to imbalances on other demographic characteristics (4, 5, 20, 21). Consistent with this account, models trained on existing majority-White datasets have been found to produce racial disparities in gender classification—that is, lower accuracy for people of color, particularly for women (4, 5). By contrast, when models were trained on more racially-balanced datasets, these disparities were reduced.

However, while informative, the existing evidence is limited because it has not established whether these disparities arise from racial composition specifically or from correlated properties of the datasets being compared. This is because the datasets used in prior work differ on dimensions besides race or ethnicity, such as image quality, sample size, annotation procedures, sampling context, preprocessing, and other demographic characteristics (e.g., age). As a result, differences associated with such datasets cannot be causally attributed to racial imbalance alone. Rather, a causal attribution would require an experimental approach, in which racial balance is manipulated while other dataset features are held constant.

Furthermore, most studies on the effects of dataset imbalance have, to date, focused on gender classification (4, 5). Although some research has observed performance decrements in race classification for minority faces trained on White-dominant data (22), the effects of dataset imbalance on race classification—and whether these effects resemble human patterns of racial bias—have not been systematically investigated.

Face classification biases in humans and AI

If racially imbalanced training data produce disparities in model performance, to what extent do these disparities resemble common human biases in race perception? Psychology research on human face categorization has identified three well-documented patterns of racial bias: the *cross-race effect*, *perceptual hypodescent*, and the *gendered-race effect*. Each captures a distinct pattern of categorization and therefore provides a social psychological benchmark for evaluating the effects of training data composition on AI face classification.

Cross-race effect. The cross-race effect (or *other-race effect*) describes the tendency for people to recognize ingroup faces more accurately than outgroup faces (23, 24)—an effect attributed to both perceptual expertise and preferences for ingroup members (25, 26). In AI systems, a cross-race pattern has been observed for facial recognition: models developed in Western countries performed more accurately for White faces than East Asian faces, whereas the opposite pattern was found for models developed in East Asian countries (27). In the context of race classification, existing studies suggest a pattern analogous to the cross-race effect (22), such that the use of majority-White training data would lead models to classify White faces more accurately than faces from underrepresented racial groups.

Perceptual hypodescent. Human racial categorization also exhibits a pattern of hypodescent: the tendency to categorize people of mixed racial ancestry according to their lower-status racial group (28). For example, White Americans tend to perceive Black–White biracial faces as Black rather than White (29–31)—a pattern linked to intergroup motives for power and resources (32, 33). Social psychological research has identified two expressions of hypodescent in humans: a classification bias, in which White participants are more likely to classify a biracial face as Black than as White (29, 33), and a representation bias, whereby participants' mental image of a Black-White biracial face is more similar to a Black than White face (34, 35). In the context of AI, research using a visual semantic model (Contrastive Language Image Pretraining; CLIP) to assign racial labels to face images found that biracial faces tended to be labeled according to their lower-status identity (36). In terms of face classification, a computational analogue would be a systematic classification of biracial faces as belonging to the lower-status racial group. Testing

this possibility requires moving beyond monoracial classification to the classification of biracial faces that vary systematically in racial composition.

Gendered-race effect. Race also shapes human gender categorization through the gendered-race effect—an intersectional expression of racial stereotypes in gender perception (37). For example, consistent with US stereotypes of Black people as aggressive and dominant and East Asian people as docile and submissive (38, 39), White American perceivers are more likely to categorize gender-ambiguous Black faces as male and gender-ambiguous East Asian faces as female (37, 40). In AI face classification, a corresponding pattern would be reflected in an interaction between race and gender in gender classification accuracy: specifically, greater accuracy for Black men than Black women, but greater accuracy for Asian women than Asian men. This pattern would expand on previously-reported performance decrements for women of color (4) to place it within a broader pattern of racial stereotype effects on gender classification.

Together, these three bias benchmarks from psychological research capture distinct dimensions of social categorization: differences in monoracial classification accuracy across racial groups, the categorization of biracial identities, and interactive effects of race and gender on gender classification. Although AI classification systems might not reproduce these human biases exactly, an assessment of where their performance mimics or diverges from patterns of human social perception can help to identify specific disparity patterns and inform their potential influence on human users.

Results

The present research tested whether racial composition in training data causally shapes face classification performance and whether the resulting disparities converge with or diverge from established patterns of human social bias. In each of three studies, we manipulated the racial composition of the training data for White faces relative to Black, East Asian, and Indian faces, using 27,040 randomly-drawn images and corresponding labels from FairFace (5). From this sample, we created four separate training datasets, each comprising 16,000 images, which varied systematically in racial composition while holding gender balance and proportion of the three minority groups constant within each condition (Figure 1). This design allowed us to compare an extreme White-dominance condition (94% White)—designed to approximate publicly-available datasets shown previously to produce biased classification (4, 5)—to conditions with progressively increasing racial balance (75%, 50%, and 25% White, i.e., fully-balanced).

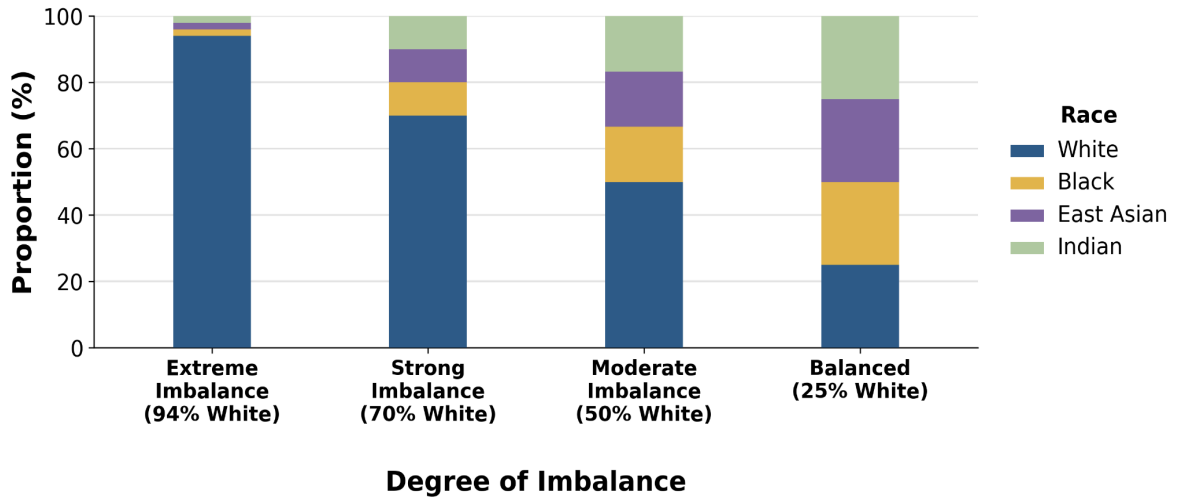


Figure 1. *Training Data Sampling Across Four Conditions with Increasing Racial Balance.* Each training dataset contained 16,000 images. Gender was balanced within each racial group (50% female, 50% male).

To build classification models, we used a Vision Transformer (ViT) pretrained on ImageNet-21k and fine-tuned it end-to-end on each manipulated training dataset, replacing the final classification layer with a task-specific linear head predicting either race (Studies 1–2) or gender (Study 3). To account for variability due to sampling and stochastic optimization, each condition was trained with five independent random seeds. For each seed, a new training dataset was generated according to the condition-specific demographic constraints (balanced versus imbalanced race distributions), ensuring that differences between conditions reflected controlled changes in data composition rather than idiosyncrasies of individual images. All models were evaluated on a balanced held-out test set containing 6,024 face images with equal numbers of images for each Race × Gender combination, such that any differences in accuracy could not be attributed to subgroup frequencies in the test data (see Methods).

Study 1

In Study 1, we tested the effect of dataset diversity on race classification accuracy (proportion of correct race predictions out of all race predictions), using the experimental manipulation and procedures for model training and evaluation described above. A seed-blocked factorial ANOVA produced significant main effects of both race, $F(3, 60) = 109.67$, $p < .001$, and training balance condition, $F(3, 60) = 62.63$, $p < .001$, indicating higher overall accuracy for White faces and for models trained on balanced datasets. However, these effects were qualified by the Condition × Race interaction, $F(9, 60) = 21.17$, $p < .001$ (Figure 2), indicating that race classification disparities depended on dataset balance.

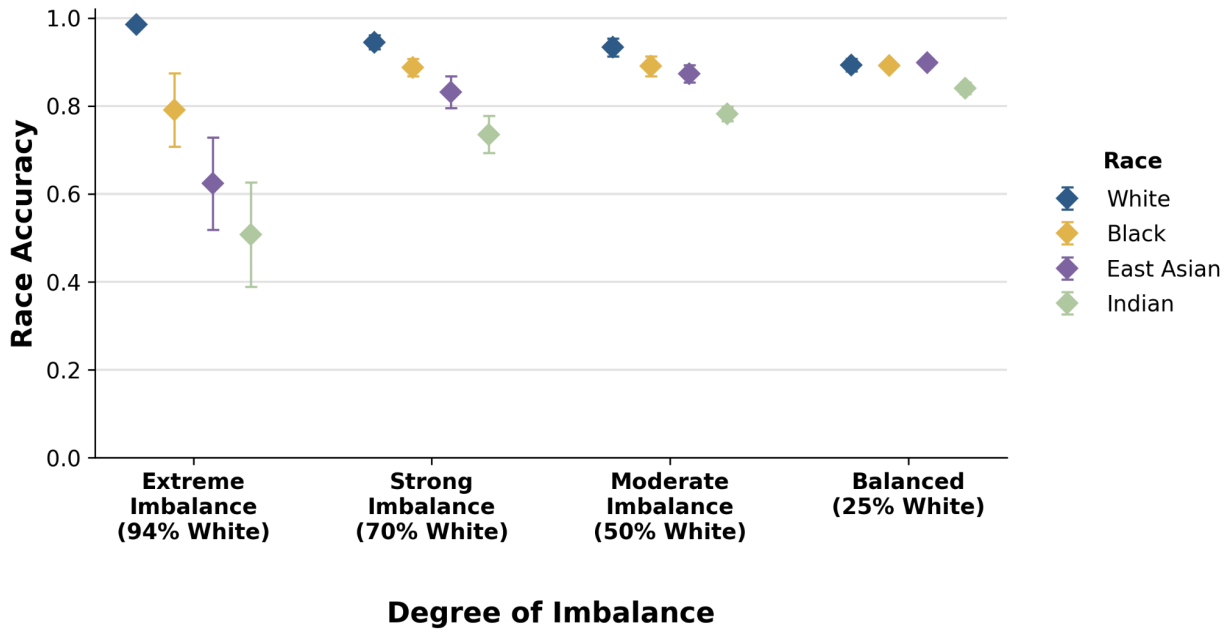


Figure 2. *Classification accuracy by training condition and race category.* Diamonds indicate mean race classification accuracy averaged across five independently trained model instantiations per condition (five random initializations) on a fixed, race-balanced evaluation set. Error bars denote 95% confidence intervals across training runs (α = .05).

To characterize this interaction, we tested simple effects of race within each condition (Table 1). The analyses revealed a large classification accuracy disparity in the extreme imbalance condition, which incrementally decreased as datasets became more balanced. In the fully-balanced condition, this disparity was fully mitigated among White, Black, and East Asian faces, with a relatively small disparity remaining for Indian faces. Notably, classification accuracy was consistently lowest for Indian faces, potentially suggesting greater variability in the images or annotator labels, among other possible factors.

In addition to reducing mean classification accuracy for underrepresented racial groups, training data imbalance was associated with greater variability across independently trained models. In the balanced condition, variability was small and comparable across racial groups (White: *SD* = .011; Black: *SD* = .007; East Asian: *SD* = .008; Indian: *SD* = .011). Under extreme White dominance, variability remained low for White faces *(SD* = .005) but increased substantially for Black (*SD* = .067), East Asian (*SD* = .084), and Indian faces (*SD* = .096). The intermediate conditions showed a graded pattern: *SD*s for underrepresented groups ranged from .013 to .018 under moderate White dominance and from .016 to .034 under strong White dominance. This

greater variability may have reflected both the smaller number of available training images and variability in the images randomly selected for each seed. This pattern suggests an effect of imbalance on model stability, in addition to classification accuracy, particularly when imbalance creates categories with relatively scarce data.

These results provide experimental evidence that training data composition causally shapes racial disparities in face classification. Models trained on extreme White-dominant datasets classified White faces more accurately than faces from underrepresented racial groups, analogous to the human cross-race effect. These disparities decreased as training datasets became more racially balanced and were largely eliminated.

Table 1

Mean Classification Accuracy Simple Effects for Race within Condition

Condition	White	Black	East Asian	Indian	*F*(3, 60)
Extreme White Dominance	.986[a]	.791[b]	.624[c]	.507[c]	32.29***
Strong White Dominance	.945[a]	.888[b]	.832[c]	.735[d]	51.78***
Moderate White Dominance	.933[a]	.891[b]	.873[b]	.782[c]	59.59***
Balanced	.893[a]	.892[a]	.899[a]	.840[b]	39.22***

Note. Means within a row that do not share a common subscript differ significantly at $p < .05$ (Tukey HSD). ***$p < .001$.

Study 2

In Study 2, we asked whether dataset diversity affects a model's classification of biracial faces—that is, whether racial imbalance alters the threshold for race classification under conditions of ambiguity. To this end, we created a controlled set of biracial test images to submit to models trained in Study 1. Test faces (N = 4,180, see Method) were morphs of real Black and White faces, representing 380 unique spectra along which faces varied parametrically in their proportion of Black-White composition on an 11-point continuum, from 0% Black (100% White) to 100% Black (0% White) (Figure 3). By assessing a model's Black versus White classification estimate for each variant within an individual spectrum, we could precisely quantify the effect of dataset imbalance on a model's Black-White classification threshold.



Figure 3: Sample face morph image spectrum, ranging from 0% Black (100% White) to 100% Black (0% White) composition in 10% increments.

Test images from each morph spectrum were submitted independently to each of the four models created in Study 1, representing the four conditions of training dataset imbalance. For each condition, across test images, classification probabilities for Black and White categories were obtained and used to compute each model's point of subjective equality (PSE)—the point along the morph spectrum at which a face is equally likely to be classified as Black or White. A PSE of 0.50 corresponds to an unbiased threshold at the objective midpoint of the continuum (morph faces with 50% White and 50% Black content are classified as equally likely to be Black versus White). Values below 0.50 indicate that faces containing less than 50% Black content (i.e., have greater White content than Black content) are equally likely to be classified as Black than as White—a pattern consistent with the perceptual hypodescent effect seen in humans (33, 41). By contrast, values above 0.50 indicate that faces containing more than 50% Black content (i.e., have greater Black content than White content) are equally likely to be classified as White than as Black—a pattern opposite to the human hypodescent bias.

To estimate condition-level PSEs, morph spectra were resampled with replacement over 1,000 bootstrap iterations. A PSE was recalculated for each resampled dataset. We report the mean of these bootstrap PSE estimates and percentile-based 95% confidence intervals, which reflect variation across morph spectra while keeping the trained models fixed.

Results indicated that race-classification thresholds differed by training set imbalance. A repeated-measures ANOVA of per-spectrum PSE scores revealed a significant main effect of training balance condition, $F(3, 1137) = 1102.29$, $p < .001$. All pairwise comparisons between conditions were significant, $ps < .001$ (Tukey's HSD corrected). However, the resulting pattern diverged from human hypodescent: in the extreme White-dominance condition, the mean PSE was .77 (95% CI [.76,.78])—a threshold substantially above 50%—indicating that a face approximately 77% Black content would have an equal chance of being classified as either Black or White. The PSE was significantly lower in the strong White-dominance condition (M=.66, 95% CI [.66,.67]), the moderate White-dominance condition (M=.69, 95% CI [.68,.69]), and the balanced condition (M=.64, 95% CI [.63,.65]). Although the balanced model's threshold was closer to .50 than other conditions, it was nonetheless significantly above the point of compositional parity.

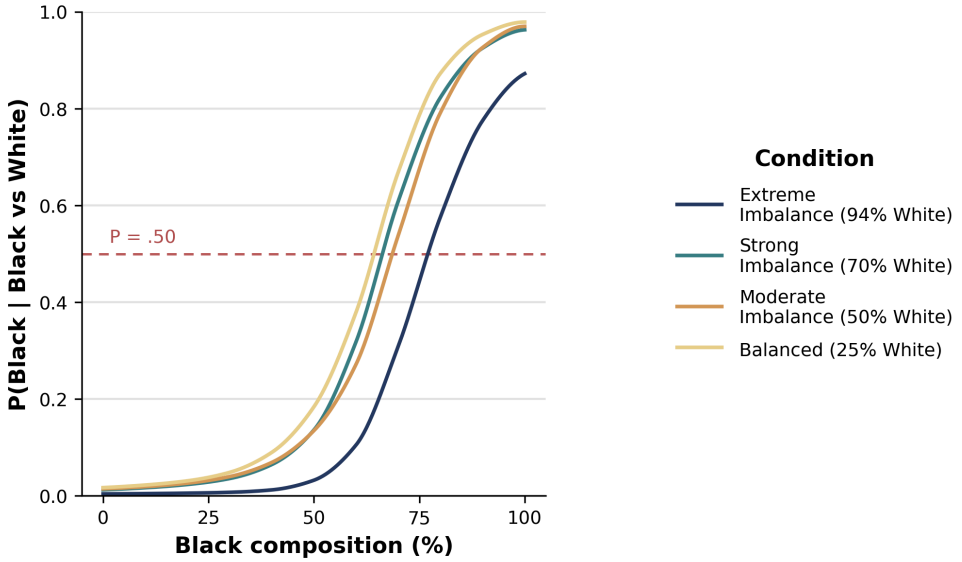


Figure 4. *Probability of Black Classification Across Morph Continua by Training Condition.* Classification curves depict the probability of assigning a face to the Black category as a function of Black composition along morph continua (from 0% Black to 100% Black), shown separately for each training condition. The point of subjective equality (PSE) corresponds to the point at which faces are equally likely to be classified as Black or White.

These results demonstrate that models trained on datasets with greater White dominance can result in a Black-White race classification threshold that requires greater Black than White racial composition. This pattern contrasts with the hypodescent classification effect observed in humans and suggests that AI systems would be more likely to identify biracial faces as White than as Black.

It is notable, however, that hypodescent in humans is also expressed in mental representations of biracial faces as being more Black than White (34, 35). An implication of the model classification pattern is that it suggests that White-dominant models—particularly the extreme model—represent biracial faces as more Black in composition than biracial faces identified by the balanced model.

To explore this possibility, we selected the 50 images closest to the classification boundary (i.e., identified as 50:50 biracial) separately for the extreme White-dominance and balanced models. As expected, biracial faces selected from the extreme White-dominance model included more Black content (M = 79.6%, Md = 80, SD = 11.24) than those selected from the balanced model (M = 65.0%, Md = 70, SD = 9.31), Mann–Whitney U = 404.5, $p < .001$. To compare differences in skin tone directly, we quantified the mean face luminance for these images. Mean image

luminance was lower for boundary images selected for the extreme White-dominance model (M = 107.72, SD = 10.81) than for the balanced model (M = 117.27, SD = 12.28), a difference of 9.55 luminance units, t = 4.13, p < .001, d = 0.83. To illustrate this effect, Figure 5 displays average morphed images of the boundary images selected for each model.

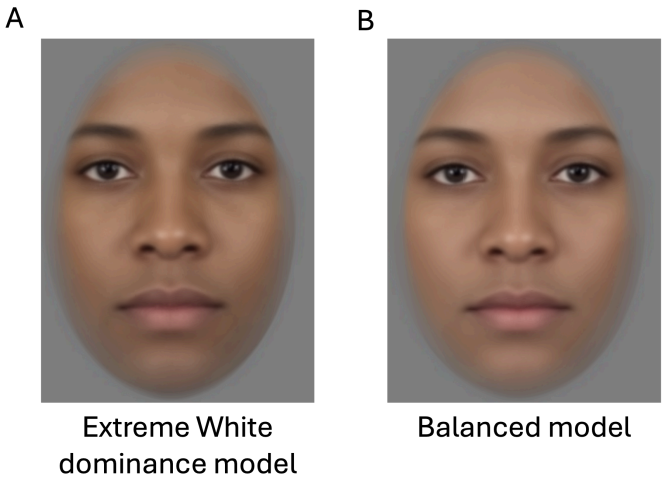


Figure 5. Composite images of faces closest to the classification boundary for models trained on (a) the extreme White-dominance dataset (94% White) and (b) the balanced dataset.

In summary, Study 2 shows that training data composition systematically shifts a model's biracial classification threshold. All four conditions yielded PSEs substantially above .5, indicating a classification boundary requiring greater Black than White content—a pattern that contrasts with human hypodescent effects in race categorization. However, this threshold was higher for models trained on imbalanced than balanced datasets. Additionally, as a consequence of these model threshold differences, White-dominant models appeared to represent biracial faces with greater Black than White content than the balanced model—a pattern that converges with hypodescent in humans' mental representations of biracial faces.

Study 3

In Study 3, we asked whether a racial imbalance in training data affects gender classification—that is, whether models exhibit a gendered-race effect, such that gender classification performance varies by race in a manner consistent with racial stereotypes. The same data and model architecture were used as in Study 1, except that we trained classification models to predict binary gender (female, male).

A factorial ANOVA with condition, race, and gender as fixed factors, and seed included as a blocking factor, produced significant main effects of condition, $F(3, 124) = 7.96$, $p < .001$, race, $F(3, 124) = 355.98$, $p < .001$, and gender, $F(1, 124) = 7.89$, $p = .006$ (Table S1 in SI). Of these, the main effect of condition indicated that overall gender accuracy improved as models became more balanced. Importantly, these main effects were qualified by a significant Race × Gender interaction, $F(3, 124) = 66.02$, $p < .001$, such that gender classification performance varied as a function of race. A significant three-way interaction, $F(9, 124) = 2.14$, $p = .030$, further suggested that the effect of race on gender classification varied by the degree of dataset imbalance (Figure 6; see Table S2 in SI for effects within condition).

To conduct a more focused test of the gendered-race effect, we reran the ANOVA while including only Black and East Asian faces. This analysis revealed a significant Race × Gender interaction, $F(1, 60) = 119.25$, $p < .001$. Simple-effect analyses indicated the gendered-race pattern: for Black faces, males ($M = .911$, $SD = .013$) were classified more accurately than females ($M = .869$, $SD = .009$), $t(60) = 5.31$, $p < .001$, whereas the reverse was found for East Asian faces, such that females ($M = .944$, $SD = .009$) were classified more accurately than males ($M = .921$, $SD = .009$), $t(60) = 2.91$, $p = .005$. However, this gendered-race pattern was not significantly moderated by dataset imbalance (three-way interaction: $F(3, 60) = 2.16$, $p = .102$).

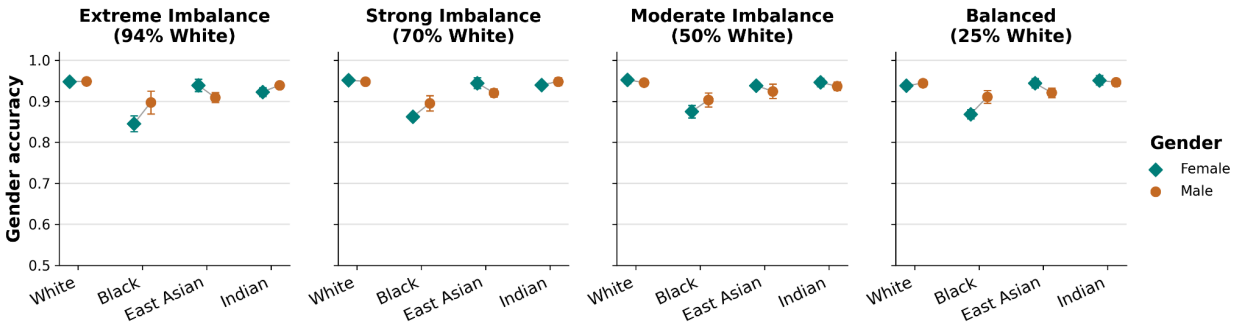


Figure 6. *Gender-classification accuracy across training conditions and Race × Gender groups.* Diamonds and circles represent mean classification accuracy across five random seeds, separately for female and male faces. Error bars represent 95% confidence intervals around the means.

Study 3 therefore revealed a prominent gendered-race effect across all face classification models. Furthermore, gender classification models did not just perform worse for underrepresented racial groups or for a specific intersectional category (e.g., Black women); rather, gender classification performance showed a pattern of bias consistent with US racial stereotypes that associate Black people with more masculine traits and East Asian people with

more feminine traits. Unlike the race classification effects in Studies 1 and 2, the use of racially-balanced training data did not significantly reduce the gendered-race effect in face classification performance.

Discussion

We investigated whether racial composition in AI training data shapes face classification patterns and whether such patterns resemble known human biases in social categorization. By manipulating dataset racial balance while controlling for image source, training set size, gender balance, and model parameters, we examined computational analogues of three well-documented patterns of human racial bias in face categorization—the cross-race, hypodescent, and gendered-race effects (42). We found that training data composition can causally alter the structure of classification behavior, producing patterns that both converge and diverge with common biases in human social perception.

Study 1 showed that models trained on datasets with disproportionately White faces, like many publicly-available datasets, performed much more accurately for White faces than Black, East Asian, and Indian faces—a pattern akin to the human cross-race effect. However, these disparities decreased progressively with greater racial balance and were largely eliminated in the fully-balanced condition. These results provide experimental evidence that racial imbalances in training data can causally produce disparities in race classification.

Study 2 extended this approach to the classification of racially ambiguous faces. Dataset imbalance again impacted performance, shifting the decision boundary for racial classification, but in a direction opposite to the bias observed in White participants. That is, whereas human perceivers tend to classify Black-White biracial faces as Black, the models here tended to classify biracial faces as White—an effect that was strongest for the extreme White-dominant model. However, as a consequence of these classification thresholds, imbalanced models appeared to represent biracial faces as having greater Black content (i.e., appeared more Black) than the balanced model—a pattern mimicking a different human expression of hypodescent whereby mental representations of biracial faces are more Black than White (34, 35). These findings suggest that depending on how an AI face classification system is used, its results could parallel or contrast with human hypodescent bias.

Study 3 examined the effect of racial imbalance on gender classification. Across conditions, models more accurately classified the gender of Black male faces than Black female faces, and of East Asian female faces than East Asian male faces—a specific interaction pattern resembling

the gendered-race effect in humans in which racial stereotypes influence gender categorization (37, 40). This finding suggests that prior observations of diminished gender classification accuracy for Black women may reflect a broader intersectional pattern involving multiple racial groups (4). Although racial balancing improved overall gender classification accuracy, it did not significantly alter the specific gendered-race pattern observed for Black and East Asian faces. Thus, the origins of this pattern remain unclear, but may be due to annotator bias in the labeling of training images, systematic features of the images themselves, or other aspects of model training.

Together, these studies reveal both the value and limits of using human social cognition as a framework for understanding algorithmic bias. The cross-race and gendered-race patterns in our models resembled established human biases, whereas the biracial classification thresholds diverged from human perceptual hypodescent. These similarities and differences suggest that examining AI outputs through the lens of human psychology can reveal distinct forms of disparity that may be overlooked by aggregate measures of model accuracy. At the same time, similarities between human and AI classification patterns should not be taken as evidence of shared underlying mechanisms. Human racial categorization is shaped by factors such as perceptual experience, stereotypes, and intergroup motivation, whereas AI classification reflects statistical regularities in training data and the optimization processes used to fit model parameters.

These findings also raise questions about how algorithmic disparities might interact with human judgment. Existing psychological research has demonstrated that people are more receptive to information that confirms their existing expectations and stereotypes, influencing consequential judgments such as eyewitness evaluations and assessments of job candidates (24, 43, 44). Research on humans' use of AI finds similar patterns of confirmation bias (9, 45, 46). Thus, if AI systems produce outputs that align with stereotypical perceptions and expectancies, users may be less likely to detect or correct for these biases. Though not the aim of the present work, identifying the effect of these bias patterns on human decision making is an important next step toward determining its impact on AI-based discrimination.

Regarding bias mitigation, our studies suggest that balanced training data can reduce some forms of algorithmic bias but not others. For AI systems designed for single-category race classification, as in Study 1, balanced race representation may effectively mitigate bias. But when the relationship between racial composition of training data and the classification objective was less direct—such as when classifying mixed-race identities or gender in Studies 2 and 3—dataset balance did not eliminate disparities. As such, additional approaches warrant investigation, including those that target model training procedures, such as efforts to prevent

hidden correlations between social categories represented in training data, as well as psychology-based interventions, such as procedures to improve accuracy and reduce bias in human annotation (6, 47–50).

Our findings highlight a broader risk of systems that impose rigid categorical identities on individuals. Classifiers that assign people to a single race or gender can misclassify those who do not fit neatly into one category (51). Such systems may also reproduce and reinforce the categorical frameworks through which social differences are constructed and reinforced, potentially contributing to discrimination and inequality (52). Future research should therefore examine not only how to improve classification within existing frameworks, but also whether systems that accommodate multidimensional or uncertain identities can reduce the harms associated with rigid categorization.

Our conclusions should be interpreted in light of several limitations. First, we examined White-majority training datasets, given the prevalence of White-dominant datasets and their relevance to racial disparities in AI. However, bias-related effects may differ under other configurations of demographic imbalance. Second, the gendered-race pattern we observed concerns Black and East Asian faces and their associated stereotypes in a US cultural context, yet different patterns may emerge in other cultural settings. Third, while our experimental design isolated effects of training data composition within our model training pipeline, other factors—such as FairFace labels, ImageNet pretraining, image preprocessing, sampling procedures, and model architecture—could also affect classification outcomes. Lastly, our design does not disentangle the effects of racial group proportions from the absolute number of minority-group images, because these features vary together as the composition of a fixed-size dataset changes. Future research could manipulate these factors independently to identify their respective contributions.

More broadly, our findings illustrate the value of an integrated computational-psychological approach to algorithmic bias, treating AI systems as products of machine learning processes embedded in human social contexts (53–55). Psychological theory provides a framework for identifying and interpreting patterns of algorithmic behavior that may not be captured by aggregate performance metrics alone (56). By examining where AI classification patterns converge with and diverge from human social biases, this approach can clarify how algorithmic disparities arise and contribute to discrimination, identify the limits of existing mitigation strategies, and guide future research on the societal impacts of AI.

Materials and Methods

Image Dataset. 27,040 unique faces were sampled from *FairFace* (5). Sampling was based on FairFace image labels, annotated by Amazon Mechanical Turk workers who, in 2020, were US-based and primarily White (57). From this pool, we constructed four training datasets of 16,000 images each, comprising unique White-, Black-, East Asian-, and Indian-labeled faces across four conditions that varied in their proportion of White faces (94%, 75%, 50%, 25%), with equal proportions of the three non-White racial groups and balanced gender within race. To include a sufficient number of images for each condition, East Asian and Southeast Asian categories were combined, and Middle Eastern, Latino, and "Other" were excluded. This sample size (16,000) was chosen because it was the largest possible sample that permitted controlled manipulations of racial imbalance while preserving gender balance within race.

Classification Models. Vision Transformer (ViT-Base/16) models (58), pretrained on ImageNet-21k, were fine-tuned end-to-end on the manipulated training datasets. The original classification layer was replaced with a newly initialized linear classification head predicting race (Studies 1 and 2) or gender (Study 3). All images were converted to RGB and processed using the pretrained model's image processor, which resized images to 224 × 224 pixels and applied the corresponding rescaling and normalization procedures. No data augmentation was used. To account for variability due to data sampling and stochastic optimization, five models were trained independently within each experimental condition, each using a separately sampled training dataset drawn from the FairFace training pool according to the condition-specific demographic constraints.

Models were trained for up to 10 epochs using the AdamW optimizer, a learning rate of 6×10^{-5}, a batch size of 64, and cross-entropy loss, with a cosine learning-rate schedule and 10% warmup. The checkpoint achieving the highest macro-F1 score on the balanced validation set was retained for subsequent analyses. Model training was conducted on the Snellius supercomputer using one NVIDIA A100 GPU, eight CPU cores, and 48 GB of system memory per training run.

Evaluation of Classification Accuracy on a Balanced Test Set (Studies 1 and 3). All models were evaluated on the same held-out test set (N = 6,024) which contained equal numbers of images for each race × gender subgroup, ensuring that differences in classification performance were not attributable to differences in the demographic composition of the evaluation data. We assessed overall classification accuracy as well as subgroup accuracy for racial (Study 1) and gender groups (Study 3). Classification accuracy was defined as the proportion of test images for which the model's predicted category matched the corresponding human annotation. In addition to overall accuracy, we calculated race-specific accuracy for Study 1 (White, Black, East Asian, and Indian) and accuracy separately for each race × gender subgroup in Study 3. Accuracy was first calculated separately for each model and subgroup and subsequently averaged across the five random seeds within each training condition.

Evaluation of Classification Threshold by Estimating Category Boundaries (Study 2). Study 2 examined classification thresholds for the racial classification of Black-White biracial faces. To create the test images that varied parametrically in Black-White racial composition, 20 Black faces (10 male, 10 female) and 20 White faces (10 male, 10 female), based on both self-identification and appearance, were selected from two popular face databases: the Chicago Face Database (59) and the Face Research Lab London Set (60). Faces were posed with a neutral expression and direct gaze and cropped. Next, 760 unique "parent" faces were created by morphing every combination of same-gender White and Black face images (190 each for Black male, White male, Black female, and White female). Mixed-race morphs were then created from each Black-White combination of parent faces within gender at 10% increments, each producing 11 face images that varied in racial composition from 100% Black to 100% White. This process produced 190 unique male and 190 unique female spectra, comprising 4180 unique test faces. Each face morph image appeared on a gray background, cropped, resized, and edited, such that the final image excluded hairstyles, necks, and ears.

Using this test set, we estimated category boundaries along controlled Black–White face-morph spectra for each Study 1 training condition. Each of the five random-seed models within a condition classified each image along each morph spectrum. Classification probabilities were then averaged across the five models, and a PSE was estimated for each morph spectrum (33). The PSE represents the point along the morph continuum at which the averaged model predictions indicate equal probability of classification as Black or White.

Differences in PSE were assessed using a repeated-measures ANOVA, with training condition as the within-spectrum factor and morph spectrum as the repeated unit. To estimate condition-level PSEs and their uncertainty, morph spectra were resampled with replacement over 1,000 bootstrap iterations. For each iteration, the average classification curve and corresponding PSE were recalculated. Mean bootstrap PSEs and percentile-based 95% confidence intervals were then obtained from the resulting distributions.

Data and Source Code Availability. Complete data and analysis code will be available on GitHub upon publication.

References

1. N. Kumar, A. Berg, P. N. Belhumeur, S. Nayar, Describable Visual Attributes for Face Verification and Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**, 1962–1977 (2011).

2. C. Su, Y. Yan, S. Chen, H. Wang, An efficient deep neural networks training framework for robust face recognition in *2017 IEEE International Conference on Image Processing (ICIP)*, (2017), pp. 3800–3804.

3. P. Perikleous, *et al.*, How Does the Crowd Impact the Model? A Tool for Raising Awareness of Social Bias in Crowdsourced Training Data in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22., (Association for Computing Machinery, 2022), pp. 4951–4954.

4. J. Buolamwini, T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, (PMLR, 2018), pp. 77–91.

5. K. Karkkainen, J. Joo, FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation in (2021), pp. 1548–1558.

6. R. K. Dass, N. Petersen, U. Visser, M. Omori, It's Not Just Black and White: Classifying Defendant Mugshots Based on the Multidimensionality of Race and Ethnicity in *2020 17th Conference on Computer and Robot Vision (CRV)*, (2020), pp. 238–245.

7. V. Dodd, UK police use of facial recognition technology a failure, says report. *The Guardian* (2018).

8. K. Hill, Wrongfully Accused by an Algorithm. *The New York Times* (2020).

9. A. Bonezzi, M. Ostinelli, Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied* **27**, 447–459 (2021).

10. G. M. Johnson, Algorithmic bias: on the implicit biases of social technology. *Synthese* **198**, 9941–9961 (2021).

11. Z. Chen, Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanit Soc Sci Commun* **10**, 1–12 (2023).

12. B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, A. K. Jain, Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security* **7**, 1789–1801 (2012).

13. H. F. Menezes, A. S. C. Ferreira, E. T. Pereira, H. M. Gomes, Bias and Fairness in Face Detection in *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, (2021), pp. 247–254.

14. A. J. O'Toole, P. J. Phillips, X. An, J. Dunlop, Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing* **30**, 169–176 (2012).

15. S. Yucer, F. Tektas, N. Al Moubayed, T. Breckon, Racial Bias within Face Recognition: A Survey. *ACM Comput. Surv.* **57**, 105:1-105:39 (2024).

16. P. Barlas, K. Kyriakou, O. Guest, S. Kleanthous, J. Otterbacher, To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proc. ACM Hum.-Comput. Interact.* **4**, 232:1-232:31 (2021).

17. A. Krishnan, A. Almadan, A. Rattani, Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, (2020), pp. 1028–1035.

18. V. Muthukumar, *et al.*, Color-Theoretic Experiments to Understand Unequal Gender Classification Accuracy From Face Images in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2019), pp. 2286–2295.

19. G. Balakrishnan, Y. Xiong, W. Xia, P. Perona, Towards Causal Benchmarking of Bias in Face Analysis Algorithms in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm, Eds. (Springer International Publishing, 2020), pp. 547–563.

20. A. Mandal, S. Leavy, S. Little, Dataset Diversity: Measuring and Mitigating Geographical Bias in Image Search and Retrieval in *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing*, Trustworthy AI'21., (Association for Computing Machinery, 2021), pp. 19–25.

21. A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana, SensitiveNets: Learning Agnostic Representations with Application to Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 2158–2164 (2021).

22. K. Kärkkäinen, J. Joo, FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. [Preprint] (2019). Available at: http://arxiv.org/abs/1908.04913 [Accessed 3 November 2026].

23. K. Hugenberg, S. G. Young, M. J. Bernstein, D. F. Sacco, The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review* **117**, 1168–1187 (2010).

24. C. A. Meissner, J. C. Brigham, Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law* **7**, 3–35 (2001).

25. M. J. Bernstein, S. G. Young, K. Hugenberg, The Cross-Category Effect: Mere Social Categorization Is Sufficient to Elicit an Own-Group Bias in Face Recognition. *Psychol Sci* **18**, 706–712 (2007).

26. S. L. Sporer, Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law* **7**, 36–97 (2001).

27. P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, A. J. O'Toole, An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept.* **8**, 14:1-14:11 (2011).

28. C. B. Hickman, The devil and the one drop rule: Racial categories, African Americans, and the U.S. Census. *Michigan Law Review* **95**, 1161 (1997).

29. A. K. Ho, J. Sidanius, D. T. Levin, M. R. Banaji, Evidence for hypodescent and racial hierarchy in the categorization and perception of biracial individuals. *Journal of personality and social psychology* **100**, 492 (2011).

30. D. Peery, G. V. Bodenhausen, Black + white = black: hypodescent in reflexive categorization of racially ambiguous faces. *Psychol Sci* **19**, 973–977 (2008).

31. D. M. Young, D. T. Sanchez, K. Pauker, S. E. Gaither, A Meta-Analytic Review of Hypodescent Patterns in Categorizing Multiracial and Racially Ambiguous Targets. *Pers Soc Psychol Bull* **47**, 705–727 (2021).

32. A. K. Ho, N. S. Kteily, J. M. Chen, Introducing the Sociopolitical Motive × Intergroup Threat Model to Understand How Monoracial Perceivers' Sociopolitical Motives Influence Their Categorization of Multiracial People. *Pers Soc Psychol Rev* **24**, 260–286 (2020).

33. A. R. Krosch, D. M. Amodio, Economic scarcity alters the perception of race. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 9079–9084 (2014).

34. E. M. Caruso, N. L. Mead, E. Balcetis, Political partisanship influences perception of biracial candidates' skin tone. *Proceedings of the National Academy of Sciences* **106**, 20168–20173 (2009).

35. A. Oliver, R. E. Tracy, S. G. Young, D. A. Wout, Black + White = Prototypically Black: Visualizing Black and White People's Mental Representations of Black–White Biracial People. *Pers Soc Psychol Bull* **50**, 1113–1127 (2024).

36. R. Wolfe, M. R. Banaji, A. Caliskan, Evidence for Hypodescent in Visual Semantic AI in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22., (Association for Computing Machinery, 2022), pp. 1293–1304.

37. P. A. Goff, M. A. Thomas, M. C. Jackson, "Ain't I a Woman?": Towards an Intersectional Approach to Person Perception and Group-based Harms. *Sex Roles* **59**, 392–403 (2008).

38. A. D. Galinsky, E. V. Hall, A. J. C. Cuddy, Gendered Races: Implications for Interracial Marriage, Leadership Selection, and Athletic Participation. *Psychol Sci* **24**, 498–506 (2013).

39. J. Schug, N. P. Alt, K. C. Klauer, Gendered race prototypes: Evidence for the non-prototypicality of Asian men and Black women. *Journal of Experimental Social Psychology* **56**, 121–125 (2015).

40. K. L. Johnson, J. B. Freeman, K. Pauker, Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology* **102**, 116–131 (2012).

41. A. R. Krosch, L. Berntsen, D. M. Amodio, J. T. Jost, J. J. Van Bavel, On the ideology of hypodescent: Political conservatism predicts categorization of racially ambiguous faces as Black. *Journal of Experimental Social Psychology* **49**, 1196–1203 (2013).

42. K. Kawakami, D. M. Amodio, K. Hugenberg, "Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization" in *Advances in Experimental Social Psychology*, (Elsevier, 2017), pp. 1–80.

43. E. V. Hall, A. D. Galinsky, K. W. Phillips, Gender Profiling: A Gendered Race Perspective on Person–Position Fit. *Pers Soc Psychol Bull* **41**, 853–868 (2015).

44. J. P. Wilson, K. Hugenberg, M. J. Bernstein, The Cross-Race Effect and Eyewitness Identification: How to Improve Recognition and Reduce Decision Errors in Eyewitness Situations. *Social Issues and Policy Review* **7**, 83–113 (2013).

45. A. Bertrand, R. Belloum, J. R. Eagan, W. Maxwell, How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22., (Association for Computing Machinery, 2022), pp. 78–91.

46. C. Rastogi, *et al.*, Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* **6**, 83:1-83:22 (2022).

47. M. Georgopoulos, J. Oldfield, M. A. Nicolaou, Y. Panagakis, M. Pantic, Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer. *Int J Comput Vis* **129**, 2288–2307 (2021).

48. L. Hickman, *et al.*, Whither bias goes, I will go: An integrative, systematic review of algorithmic bias mitigation. *Journal of Applied Psychology* **110**, 979–1000 (2025).

49. M. A. McDaniel, N. S. Hartman, D. L. Whetzel, W. L. Grubb Iii, Situational Judgment Tests, Response Instructions, and Validity: A Meta-Analysis. *Personnel Psychology* **60**, 63–91 (2007).

50. M. Parmar, S. Mishra, M. Geva, C. Baral, Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. [Preprint] (2024). Available at: http://arxiv.org/abs/2205.00415 [Accessed 4 November 2025].

51. A. Yi, B. Turner, Representations and consequences of race in AI systems. *Current Opinion in Psychology* **58**, 101831 (2024).

52. M. Cikara, J. E. Martinez, N. A. Lewis, Moving beyond social categories by incorporating context in social psychological theory. *Nat Rev Psychol* **1**, 537–549 (2022).

53. S. Barocas, M. Hardt, A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press, 2023).

54. A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and Abstraction in Sociotechnical Systems in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19., (Association for Computing Machinery, 2019), pp. 59–68.

55. M. Vlasceanu, D. M. Amodio, Propagation of societal gender inequality by internet search algorithms. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2204529119 (2022).

56. G. Gigerenzer, Psychological AI: Designing Algorithms Informed by Human Psychology. *Perspect Psychol Sci* **19**, 839–848 (2024).

57. A. J. Moss, C. Rosenzweig, J. Robinson, S. N. Jaffe, L. Litman, Is it ethical to use Mechanical Turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages. *Behav Res* **55**, 4048–4067 (2023).

58. A. Dosovitskiy, *et al.*, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale in (International Conference on Learning Representations (ICLR), 2021).

59. D. S. Ma, J. Correll, B. Wittenbrink, The Chicago face database: A free stimulus set of faces and norming data. *Behav Res* **47**, 1122–1135 (2015).

60. L. DeBruine, B. Jones, Face Research Lab London Set. (2017).

Supplementary Information for

Data Diversity Effects on Algorithmic Bias:
Do Racially-Imbalanced Training Data Produce Human-Like Face Classification Biases?

Jeffrey J. Berg[1*], Anke Stoll[2*], Qianqian Cui[3], & David M. Amodio[2]

[1]New York University, [2]University of Amsterdam, and [3]University of California, Santa Barbara

Supplementary Results

Table S1
Study 3: Full Omnibus ANOVA Results (All Conditions)

Effect	*F*	*p*	*Partial η²*
Condition	F(3, 124) = 7.96	< .001	.161
Race	F(3, 124) = 355.98	< .001	.896
Gender	F(1, 124) = 7.89	.006	.060
Condition × Race	F(9, 124) = 3.04	.003	.181
Condition × Gender	F(3, 124) = 1.83	.042	.142
Race × Gender	F(3, 124) = 66.02	< .001	.615
Condition × Race × Gender	F(9, 1244) = 2.14	.030	.135
Seed (blocking factor)	F(4, 124) = 0.26	.905	.00

Table S2
Study 3: Simple Effect tests of the Race × Gender Interaction for Each Training Condition

Training condition	*Race × Gender interaction*	*p*	*Partial η²*
Extreme imbalance	F(3, 28) = 19.13	p < .001	.672
Strong imbalance	F(3, 28) = 20.22	p < .001	.684
Moderate imbalance	F(3, 28) = 9.19	p < .001	.496
Balanced	F(3, 28) = 24.89	p < .001	.727