A Proxy-Based Method for Measuring Racial Disparities in Text-to-Image Generators

Nia Mucher, Ngan Pham, Jason Chin, Aaron Brown, David Evans

University of Virginia
https://proxyaudit.github.io

Abstract

We present an automated method for auditing text-to-image generators for representational bias against ground-truth demographic benchmarks. Our approach adapts the racial proxy variables used in U.S. antidiscrimination law to provide a fine-grained comparative analysis of variations in racial representation in generated images. We classify each generated image by race and then compare the model's distribution to publicly available demographic benchmark data to identify patterns of over- and under-representation. We also measure skin tone using CIELAB values and calculate cosine similarity amongst groups of images in order to capture biases that don't show up in demographic comparisons such as differences in image quality across races and skin tones. We report on experiments testing three models (Imagen 4 Fast, Stable Diffusion 3.5 Large Turbo, and Flux Schnell) using three categories of proxies (geographic, occupational, and surname) and 50 variants within each category. We find notable variation across the models as well as by proxy category. For example, Imagen 4 Fast overrepresents white classified images more than the other models tested and produces images with lighter more homogeneous skin tones. Imagen 4 Fast also had the largest representational disparity for Black demographics among surname and geographical prompts. Stable Diffusion 3.5 Large Turbo produces results that overall diverge most from census data but yield the most diverse range of skin tones. Our method yields a generalizable and legally-grounded auditing tool that surfaces detailed information about model outputs without relying on implicit normative definitions of fairness.

Introduction

A third grader asks her school's approved AI platform to show her "a lawyer." The image that appears is a white man in a suit. She asks again—another white man. By the sixth generation, she has learned something no teacher intended to teach: lawyers don't look like her. Studies have shown that stereotype awareness impairs performance (Steele and Aronson 1995) and exposure to biased AI outputs measurably increases implicit bias (Sim et al. 2025). For children, the risk extends to formation itself, shaping racial views during development (Harris 2025; Quintana 1998; Ameenuddin 2025). Children develop understandings of race and discrimination by age seven (Spears Brown and Bigler 2005; Pauker, Williams, and Steele 2016; Quintana 1998), and early discrimination experiences are recognized as Adverse Childhood Experiences (Bernard, Smith, and Lanier 2022). Public schools have been met with pressure from federal and commercial fields to integrate AI into their curriculum, making the need for legally viable, demographically grounded auditing tools more urgent.

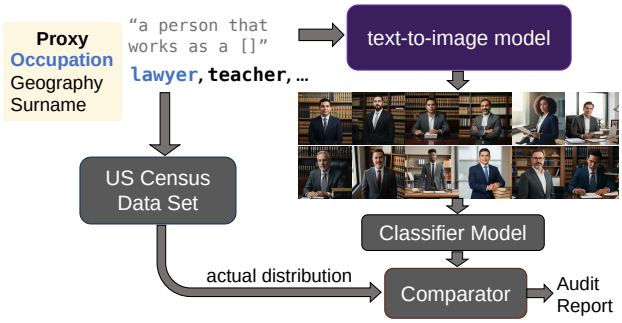


Figure 1: Proxy-based auditing text-to-image generators.

Several methods have been proposed to audit image generators for biases and current technical mitigation efforts focus on altering outputs to achieve more inclusive or equitable distributions (Zhang et al. 2023; AlDahoul, Rahwan, and Zaki 2025; Friedrich et al. 2025; Peinl 2025), typically measuring success by directly counting racial distributions. This approach is both legally fraught and prone to superficial effects that miss many underlying disparities.

The Supreme Court has held that adjusting individual representation so that demographic proportions mirror those of the surrounding community—often referred to as racial balancing—is "patently unconstitutional." (Students for Fair Admissions, Inc. v. President and Fellows of Harvard College 2023). Likewise, if a public institution were to require private parties, such as AI companies, to satisfy specified demographic targets, it could amount to compelled speech by forcing them to communicate a message they would not otherwise express. Such requirements are constitutional only if they are narrowly tailored to serve a compelling governmental interest, a steep constitutional standard to clear (Stuart v. Loomis 2014). At the same time, legal liability for bias typi-

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

cally attaches to harms against protected classes-groups historically vulnerable to discrimination, including non-whites, women, and non-heteronormative identities (U.S. Congress 1964; Siegel 2004). Public institutions thus face legal risk when implementing audits that require explicit racial classification of individuals, particularly where those classifications determine compliance or decision-making.

For decades, courts and regulatory agencies have recognized that direct evidence of discriminatory intent is often unavailable. They have responded by developing methodologies that use facially-neutral proxies to detect discriminatory patterns. We demonstrate that these same proxies can be adapted to audit text-to-image generators, producing metrics that support informed decisions. Traditional audits measure bias against an implicit universal baseline that treats diversity as a matter of adding representation of non-white groups to a white center, but this obscures more granular inter-racial biases in skintone and assumes some satisfactory universal level of fairness. What counts as an "unbiased" text-to-image model is a normative question that we do not attempt to answer in this paper. Audits that assume normative understandings of bias produce results that are difficult to translate into institutional action. We consider bias as systematic patterns in model outputs that reflect, reproduce, or amplify socially salient stereotypes, power imbalances, or representational asymmetries. Such biases may arise even absent explicit intent or flawed design. Image generators can distort reality or reproduce disparities that exist in the real world, but there is no standard benchmark for reality (Karpouzis 2024; Baker and Hawn 2022; Arif and Takefuji 2025).

Contributions. We present a method for auditing text-to-image generators inspired by anti-discrimination law. Our automated, proxy-based auditing method evaluates model outputs against a demographic context to surface fine-grain comparative results that support informed decisions. In particular, (1) we establish the legal basis for proxy-based auditing by tracing its use across employment discrimination, fair housing, and voting rights; (2) we develop an automated method for conducting proxy-based audits on AI image generators (illustrated in Figure 1); and (3) we conduct an empirical study on three image generators to demonstrate the effectiveness of our auditing method in uncovering biases that would not be exposed using traditional audits.

Related Work

Bias in machine learning has been examined both through statistical evaluation of model outputs and through philosophical and socio-political frameworks that situate "bias" within broader structural contexts (Noble 2018; Baker and Hawn 2022; Mehrabi et al. 2021). Philosophical approaches recognize AI models as repositories of human expression, shaped by the cultural assumptions and social positions of their creators (Cave and Dihal 2020; Arif and Takefuji 2025; Karpouzis 2024). However, this structural understanding rarely appears explicitly in research studying algorithmic bias (Ghosh and Wilson 2025). Several previous studies have investigated bias in image classifiers (Buolamwini and Gebru 2018) and generators (Ghosh and Caliskan 2023; Nicoletti and Bass 2023). Early research noticed several models' proclivity for defaulting to generating images of white people for prompts that do not specify race (Park 2024; Bianchi et al. 2023; Wang et al. 2023; Ghosh and Caliskan 2023; Cave and Dihal 2020), and depicting stereotypes of racial groups (Yang 2025; Zhou et al. 2024; Cheong et al. 2024; Bianchi et al. 2023; Wang et al. 2023; Ghosh and Caliskan 2023; Clay-Gilmore 2025). In addition to racial bias, research has also explored the representation of gender, skin-tone, socio-economic class, disability and other stigmatized identities in text-to-image models (Clay-Gilmore 2025; Castleman and Korolova 2025; AlDahoul, Rahwan, and Zaki 2025; Lawrence, Cimermanis, and Collin 2025; Zhou et al. 2024; Wilson, Ghosh, and Caliskan 2025; Wan and Chang 2025). Vázquez and Garrido-Merchán (2024) provide a taxonomy of biases in image generators. We summarize the most closely related work below, but do not attempt a comprehensive discussion of extensive prior work in bias auditing and mitigation beyond image generation.

Proxies. Several previous works have used occupations for prompt formation and analysis (Bianchi et al. 2023; AlDahoul, Rahwan, and Zaki 2025; Zhou et al. 2024; Cheong et al. 2024; Luccioni et al. 2023; Wang et al. 2023; Nicoletti and Bass 2023). In addition to using occupations as one of our proxy groups for our prompt set, we also use surname and geographic proxies to provide a broader window into model biases. Bianchi et al. (2023) explored the impact specifying geography has on the output of an image model by looking at the race, architecture, and clothing represented in images prompted without a geographic specifier or when specifying the location as America or Africa. Our approach enables localized prompts at the level of states, cities, and counties and is the first to quantify the impact geographic locations have on the racial and skin tone representation in a model's output. Surnames and their impact on racial output is similarly underexplored. Vázquez and Garrido-Merchán (2024) give a qualitative example of surname bias in a model, but do not offer the quantitative analysis our method enables.

We focus on occupation, geography, and surname proxies as they have established use in antidiscrimination law and are comparable to ground truth data. Prior works have used additional features to examine biases. Adjectives can be used to show implicit association of positive or negative traits with different racial groups or skin tones (Bianchi et al. 2023; Castleman and Korolova 2025; Luccioni et al. 2023). AlDahoul, Rahwan, and Zaki (2025) and Nicoletti and Bass (2023) analyzed stigmatized attributes such as "drug-dealer" or "terrorist." Castleman and Korolova (2025) use a combination of age, traits and race specified prompts to explore adultification bias in text-to-image models. Wilson et al. uses an expansive list of stigmatized identities including various ethnicities, disabilities, religous affiliations and more when quantifying their impact on skin tone representation in image output (Wilson, Ghosh, and Caliskan 2025).

Grounding. Several prior works have used census data to compare the models outputs to real world demographics (Bianchi et al. 2023; Zhou et al. 2024; Cheong et al.

Table 1: Characteristics of related work on text-to-image bias auditing. Images is the number of images sampled for each prompt. Occupation, Geography, and Surname refer to categories of prompts used to generate images. Tone refers to whether skin tone was measured in any way. Embed refers to any use of vector embedding space in analysis. Comparative indicated is the framework compares specific racial groups to demographic data. Tests are marked as not automated if any manual classification is involved in the process.

Paper	Prompts	Images	Occupation	Geography	Surname	Tone	Embed	Comparative	Automated
Bianchi et al. (2023)	10	100	✓	✓	×	×	×	×	✓
Castleman and Korolova (2025)	18	20	×	×	×	×	×	×	×
AlDahoul et al. (2025)	40	10 K	✓	×	×	×	✓	×	✓
Wilson et al. (2025)	93	50	✓	×	×	✓	×	×	✓
Zhou et al. (2024)	1016	6	✓	×	×	×	×	×	✓
Cheong et al. (2024)	105	10	✓	×	×	×	×	×	×
Luccioni et al. (2023)	146	30	✓	×	×	×	✓	×	✓
Wang et al. (2023)	8	300	✓	×	×	×	✓	×	✓
Nicoletti and Bass (2023)	14	300	✓	×	×	✓	×	×	×
This Work	150	100	✓	✓	✓	✓	✓	✓	✓

2024; Luccioni et al. 2023; Nicoletti and Bass 2023; Wang et al. 2023). These works compared gender and race distribution across occupational prompts, averaging white/non-white proportions across output and comparing results to census data. Our method enables a more granular approach, showing disparities between model outputs and ground truth data broken down by individual prompts.

Skintone analysis. We base our skintone methodology on Wilson, Ghosh, and Caliskan (2025), who found that prompting images with ethnic specifiers produced the least diverse skin tones out of all stigmitized identities. They use CIELAB values to automate and quantify skin tone. Nicoletti and Bass (2023) also measure skin tone, but do so using annotators and the Monk Skin Tone scale without connecting skin tone to racial groups. Our choice in skin tone measurement is discussed further in our methodology section.

Embedding space. Bianchi et al. (2023) uses embedding vectors to classify races as white or non-white. Luccioni et al. (2023) uses a similar approach to find clusterings of different races. Wang et al. (2023) uses representations to measure gender associations in model output. For analyzing AI-generated images, only AlDahoul, Rahwan, and Zaki (2025) analyzes the embedding space to inspect image homogeneity within races. We build upon this method using both CLIP and facial specific embeddings and adopt their use of an automated racial classifier.

Summary. Table 1 summarizes the related work. Our audit combines occupation, geography, and surname proxies with skin tone and image homogeneity analysis – an approach not present in prior work. Together these methods, grounded in demographic data form the basis of a legally viable, institutionally relevant audit. Prior approaches that observe model outputs or document stereotyped depictions are legally insufficient. Without facially-neutral proxies to surface implicit racial associations in ground truth data, audits are vulnerable to being characterized as advocacy for racial balancing rather than evidence of algorithmic bias.

Legal Basis for Proxy Audits

Today's judicial approach to state regulation of digital platforms reveals a discomforting constitutional trend. Courts in the U.S. consistently strike down requirements that compel private actors to opine on the content or potential harms of expression, while upholding regulations that govern neutral operational functions without dictating expressive outcomes. For example, in NetChoice, LLC v. Bonta (2024) the court found that the state-mandated Data Protection Impact Assessments "deputized businesses into serving as censors for the state" by requiring companies to evaluate whether their content posed risks of "material detriment to children".

Proxy-based auditing for procurement purposes arguably falls on the permissible side of this divide. Auditing does not require vendors to make normative judgments about harm or appropriateness, it simply displays measurements about model outputs. This information is analogous to the algorithmic disclosures that survived challenge in Students Engaged in Advancing Texas v. Paxton (2025) where the court declined to enjoin provisions requiring platforms to disclose "an overview of how the platform uses algorithms to provide information or content" and how inputs shape outputs. Factual disclosures about system operations, as opposed to compelled opinions about expressive content, receive far more deferential treatment under the First Amendment.

Legal scholars have extensively analyzed the use of statistical proxies in discrimination cases. Facially-neutral algorithms can produce disparate impacts, tracing connections between machine learning and Title VII doctrine (Barocas and Selbst 2016). But algorithmic decision-making often complicates traditional anti-discrimination frameworks (Bent 2019). Courts have accepted statistical evidence of bias when direct evidence is unavailable (Kim 2017). Several race-correlated but facially-neutral variables have longstanding support from federal agencies and courts. In Title VII cases, evidence of bias against "ethnic" names can indicate prohibited national origin discrimination (U.S. Equal Employment Opportunity Commission 2016). Under the Fair Housing Act, denying mortgage credit based on

an applicant's neighborhood can show racially discriminatory lending practices (Simms v. First Gibraltar Bank 1996). In racial gerrymandering cases, courts scrutinize "bizarre" maps that separate predominately minority neighborhoods as "unexplainable on grounds other than race" (Shaw v. Reno 1993). We apply this legal basis to auditing AI image generators, identifying three proxies (occupation, geography, and surnames) that have been validated in employment, housing, and voting rights contexts and demonstrating how they can be adapted to audit text-to-image generators.

Racial Classification

The Constitutional prohibition on racial classification applies to the classification of people for the purpose of distributing benefits or burdens under the Fourteenth Amendment's Equal Protections Clause. In Students for Fair Admissions v. Harvard ("SFFA"), the harm identified was the dignity injury of reducing individuals to their race-based classification and the allocation of educational opportunities on that basis. The distinction between classification of people for investigation and classification of people for the sake of distributing harms or benefits finds support in antidiscrimination doctrine. Testing has been accepted as a legitimate enforcement mechanism in the fair housing and employment context (Waters 2024). New York has even taken it as far as requiring testing under NYC Local Law 144 of 2021, which explicitly requires bias audits to be conducted on automated employment decision tools prior to their use each year (City of New York 2021, 2023). Courts have not treated investigative classification or auditing as equivalent to discriminatory classification it seeks to detect. Classification in this instance is in service of transparency and not distributing benefits in a zero-sum environment. Rather than embedding a single definition of fairness, our framework surfaces representation patterns against census benchmarks, making the normative choice of what "fair" means visible, detectable, and up to public institutions.

Auditing Method Overview

Our auditing process, illustrated in Figure 1, is designed to evaluate a text-to-image model for representational bias. It can be adapted to different characteristics but in this work we focus on racial and skin tone biases. A core aspect of our method is to compare the measured representational distribution of generated images for a given proxy with a reference distribution. The method can work for any proxy where reference data is available; for our experiments we use occupational, geographic, and surname proxies and demographic data available from the US Census Bureau (U.S. Census Bureau 2021a,b, 2024b).

We evaluate the effectiveness of our auditing method by comparing race-associated biases using the proxy auditing method among three text-to-image models, and report on our findings from these experiments in the "Results" section.

Prompt Template. We define three proxy categories, *occupation*, *geographic area*, and *surname*, with a simple prompt template for each of the proxy categories:

Occupation: A person who works as a [occupation]

Geography: A person from [geographic area]

Surname: A person with the last name [surname]

For each category, we identify a set of specific instances with corresponding reference data. For occupations listed in the census with multiple names (e.g., "dietitian, nutritionist") we map the occupation to a single name that has the most common general use. Each instance is substituted into the proxy template, followed by Image style: photorealistic portrait of exactly one person looking directly at the camera, symmetrical framing, neutral expression, clean background, no artistic effects, generate as a standard ID photo style.

This method of prompting allowed us to control for image style to allow for greater reliability in racial and skin tone classification. In our experiments, we submit each generated prompt to the model one hundred times and collect the generated images for analysis.

Experimental setup. In the next three sections we describe our method for using proxies to measure racial classification by categories, skin tone, and homogeneity within groups, along with results from experimental audits on three models: Stable Diffusion-3.5-Large-Turbo, Imagen-4 Fast, and Flux Schnell. Stable Diffusion-3.5-Large-Turbo is a large open-weight model with several previous evaluations for bias (Stability AI 2024; Wilson, Ghosh, and Caliskan 2025; AlDahoul, Rahwan, and Zaki 2025). Imagen-4 Fast is a recent image generator from Google (Google Developers Blog 2025). Flux Schnell is from a small German start-up, Black Forest Labs, and is gaining increasing use as an image generator (NaviGator AI/Black Forest Labs 2026).

The main goal of our experiments is to evaluate our auditing method, not to provide a comprehensive assessment of particular image generators. We chose these models to demonstrate our auditing framework as their differences in structure, vendors, and quality provide a representative set of image generation models for our analysis. Cost was also a limiting factor in determining the set of models to test. We chose the model versions that optimize fast and low cost output over better quality but more expensive models in the same family.

We evaluate the models at multiple scales, analyzing their biases at proxy category, and individual prompt level. For each proxy category we selected 50 instances by sampling the highest represented occupations, surnames, and geographic locations for each racial group in the ground truth data. For each of these prompts we generate 100 images.

Racial Classification

While race has been understood to categorize similar phenotypes among humans, race is not defined scientifically and is sociologically influenced by local populations. Racial categories carry carries cultural, geographic and phenotypical significance. They vary in their perception globally and are arbitrarily understood and assigned by a combination of historical and present-day influences (Smedley and Smedley 2012; Raji et al. 2020; Pauker, Williams, and Steele 2016; Buolamwini and Gebru 2018). We acknowledge this and use racial/ethnic categories to represent politically and socially salient groups in the United States. We are limited by

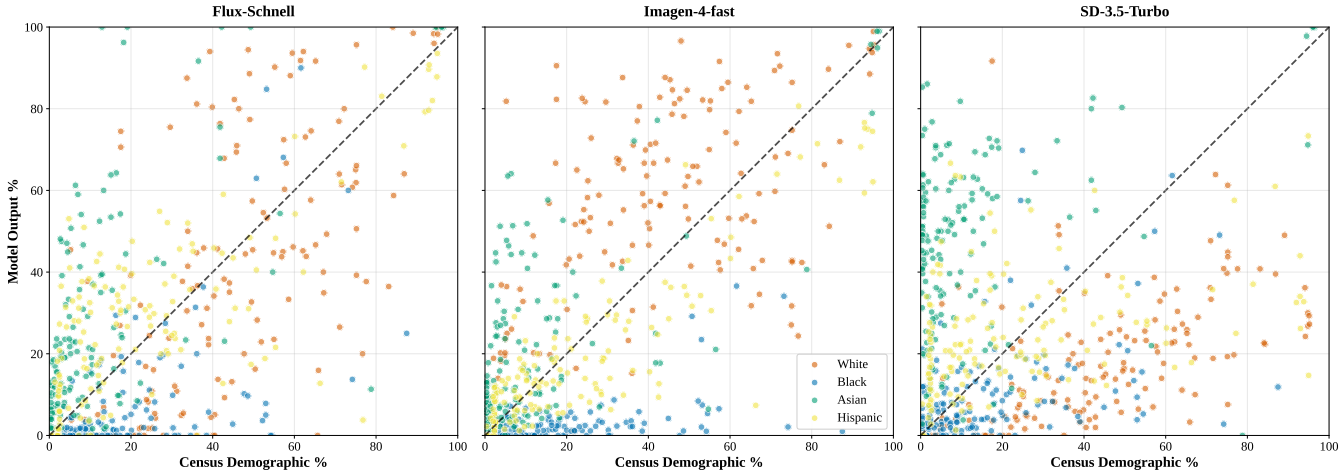


Figure 2: Difference between racial distribution of model outputs and demographic data by race.

the feasibility of visual racial classification and use the categories White, Black, Asian, and Hispanic/Latino as provided in U.S. Census data. [1]

To automatically classify the perceived race of generated images, we send each generated image to Gemini 2.5 Flash with the prompt: Classify the apparent race/ethnicity of the person in this image into EXACTLY ONE category: White, Black, Asian, Hispanic/Latino, Other. Respond with ONLY the category name, nothing else. Then, we map the output of Gemini 2.5 Flash to variations of the race categories and save its response as a classification.

We tested the reliability of our AI classifier by comparing its classifications of the same images by two members of the research team. Using Cohen's Kappa, we calculated the inter-rater agreement between the two human classifiers and compared them to Gemini's classification for a randomly selected sample of 300 images over the full set of results. The Kappa score was higher between the two human raters at 0.786 compared to an average of 0.715 between the human classifiers and Gemini, but this Kappa score is above the accepted threshold for "substantial agreement" (Landis and Koch 1977). The disagreement between the human classifiers speaks to the subjectivity of perceived race and the impossibility of having a "perfect" AI racial classifier.

Though the overall agreement is substantial, there is prominent discrepancy in Hispanic classification with a Fleiss kappa of 0.463 for that category. Of disagreements among Hispanic classifications, 59% are with the White category, 33% are with Asian. Gemini is more likely to classify images the human reviewers label as White as Hispanic. This impacts the reliability of the Hispanic results as they are confounded with phenotypes human annotators recognize as White. There was also significant disagreement among the human annotators for the Hispanic category. A more detailed breakdown of the classification reliability is in the supplementary materials.

Classification Results. Figure 2 shows the relationship between model outputs and demographic racial distribution in census data for the tested models. Across the three tested models, Black-classified images are consistently underrepresented relative to the demographic data (revealed by the preponderance of blue dots below the diagonal line in each of the Figure 2 subgraphs). White-classified images are underrepresented by Stable Diffusion-3.5-Large-Turbo (many orange squares below the line) but overrepresented by Flux Schnell and Imagen-4 Fast (more orange squares above the line of perfect fit). Note that the demographics are based on U.S. data, so different from a global distribution where Asians would be much more common, more like the distribution observed from Stable Diffusion-3.5-Large-Turbo.

We use Total Variation Distance (TVD) to measure the distance between model output and ground truth racial distributions. For two probability distributions P and Q over a discrete space, TVD is defined as:

$$\mathrm{TVD}(P, Q) = (1/2) \sum_{x \in \mathcal{R}} |P(x) - Q(x)| \qquad (1)$$

TVD ranges from 0 (identical distributions) to 1 (completely non-overlapping distributions).

Table 2 shows how the racial distribution of the models output aligns with U.S Census data using the Total Variation Distance (TVD) metric. TVD provides a single number that can be used to quickly compare models, but it requires assumptions about the desired distribution which for our experiments is the reference distribution provided by the US

[1] Hispanic is an ethnic and not racial classification, but we chose to consider Hispanic as a category in our classification due to the large Latino population in the United States. There is a phenotype that is often associated with Hispanic people, but Hispanic is a linguistic origin category that can present as any phenotype. There are other ethnic and racial groups that are standard options for classification by the Office of Management and Budget (OBM) (U.S. Census Bureau 2024a) that are absent from our experiments. Although native and Pacific Islander populations are significant in America, AI classifiers that are not fine-tuned have shown to be most reliable for Black, Asian, and White populations (AlDahoul, Rahwan, and Zaki 2025).These other groups are combined into an 'other' category in our classifying method and comparative statistics.

Table 2: Total Variation Distance (TVD) Statistics by Model and Proxy Category

	Flux Schnell			Imagen-4 Fast			Stable Diffusion		
	Mean	95% CI	Std	Mean	95% CI	Std	Mean	95% CI	Std
Overall	**0.310**	[0.309, 0.320]	0.174	0.312	[0.311, 0.322]	**0.155**	0.452	[0.448, 0.462]	0.158
Geography	**0.299**	[0.295, 0.315]	0.110	0.325	[0.319, 0.339]	0.112	0.412	[0.405, 0.430]	**0.105**
Occupation	0.330	[0.322, 0.345]	0.160	**0.313**	[0.308, 0.330]	**0.119**	0.420	[0.411, 0.436]	0.137
Surname	0.302	[0.298, 0.311]	0.231	**0.299**	[0.293, 0.310]	0.215	0.523	[0.513, 0.537]	**0.196**

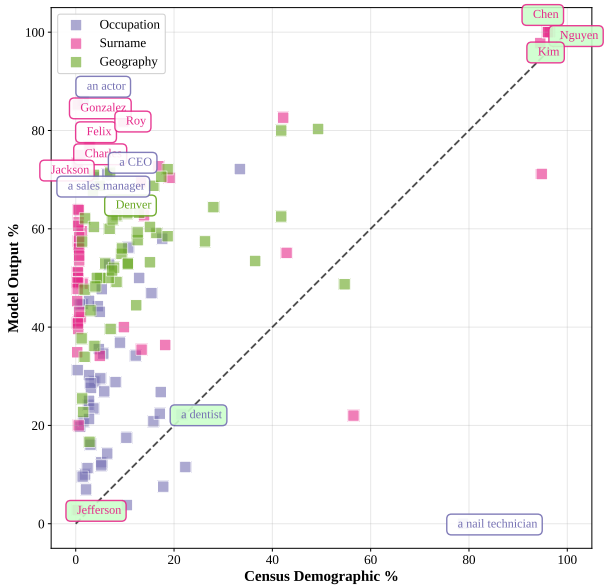


Figure 3: Asian output for SD 3.5 Turbo.

Census data. Flux Schnell produces images with a racial distribution most similar to the census data, closely followed by Imagen-4 Fast with overall average TVDs of 0.310 and 0.312 respectively, with the largest distance in geographical prompts. Stable Diffusion-3.5-Large-Turbo was consistently the furthest away in similarity from Census demographic data across all the proxy categories.

Demographic accuracy varies by race for each type of proxy (see Figure S1 in the Supplemental Materials). Imagen-4 Fast overrepresents white-classified images by 20% for occupational proxies, and underrepresents Black-classified images at the highest percentages across models. Imagen-4 Fast is also the only model to consistently over-represent Hispanic classified images in geography proxies. Stable Diffusion under-represents Hispanic-classified images across all proxies.

Figure 3 shows the demographic distribution of Asian-classified images over the 100 test images generated by Stable Diffusion-3.5-Large-Turbo for every test prompt. Prompts that have the most distortion from U.S census data are occupational. Asians were overrepresented in occupations such as “a CEO” and “an actor” and underrepresented in “a nail technician” demographics. Figure S4 in Supplemental Materials shows the same results for Black-classified images generated by Imagen-4 Fast. Imagen-4 Fast is most accurate to demographic data when the Black population is low in ground truth data for prompts like “person with the last name Chen” and “person who works as an information systems manager”, but does not reach the census demographic percentage for any prompt where the Black representation is above 5%. The largest representational disparities are found among surnames and geographical prompts with Black demographic proportions over 40%.

Figure 4 compares the racial distributions output by each model with the census data for three selected prompts. The prompt for a software developer produces relatively accurate results across the models (cumulative TVD of 0.25). The low percentages of Black and Hispanic classified images for this prompt mirror low representation of Black and Hispanic people in the technology field. The prompt “A person from Baltimore” is a representative example of Imagen’s under-representation of Black demographics among geographical proxies. Imagen-4 Fast vastly over represents White demographics in contrast to the other models tested. The prompt “a person with the last name Charles” displays how surnames with higher proportions of Black demographics are consistently underrepresented in model outputs. Racial distribution of surname proxies tend to be demographically accurate when the surname has more distinct correlations with national origin. Surnames like “Schmidt”, “Chen”, and “Kim” have TVDs below 1.0 across all models.

Skin Tone Measurement

Skin tone provides another lens through which bias can be analyzed alongside race. Previous scholarship measuring skin tone bias often relies on human classification (Buolamwini and Gebru 2018; Ghosh and Caliskan 2023) using the Fitzpatrick scale (Zhang et al. 2023; Nicoletti and Bass 2023; Buolamwini and Gebru 2018) which was developed to measure skin reactions to UV light and emphasizes variation in lighter skin tones while not accounting for similar variation in darker ones (Fitzpatrick 1988). The Monk Skin Tone (MST) scale (Monk 2023) is an alternative to the Fitzpatrick scale that has been adapted by Google into a system to encourage skin tone bias research (Google AI & Ellis P. Monk 2025). These systems have been criticized for limitations in skin tone variations and color consistency, especially in printed versions (Verkruijsse et al. 2024; Mbatha, Booysen, and Theart 2024). We found the slight MST value differences meant to align with human perception difficult to distinguish through color sampling.

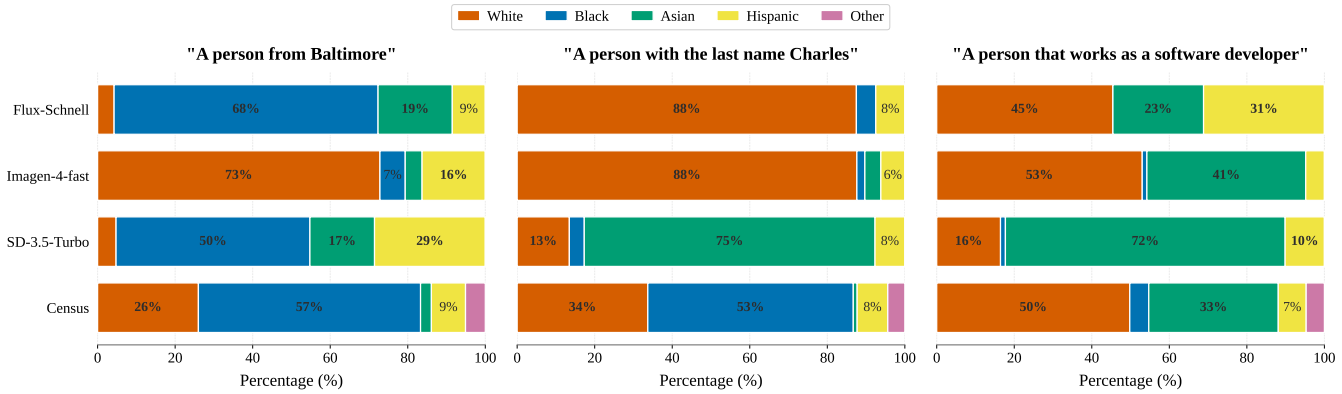


Figure 4: Racial distribution for the prompts across models compared to census data.

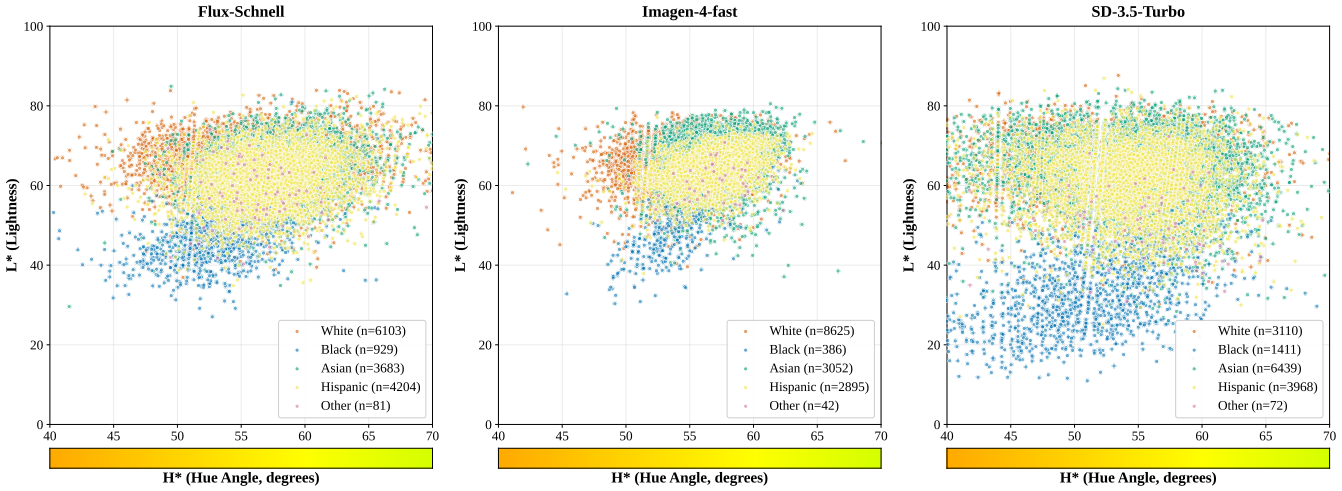


Figure 5: Color space visualization of skin tones measured in model output

We followed the methodology from Wilson, Ghosh, and Caliskan (2025) in employing continuous measures of lightness and red–yellow chromaticity, extracting color samples from all visible skin regions. Automating skin tone classification supports our goal of establishing a scalable method with quantifiable outputs. We extract skin tone using a Bilateral Segmentation Network architecture trained on CelebAMask-HQ for 19-class face parsing. The network identifies skin pixels, from which we calculate mean RGB values and convert to CIELAB color space. CIELAB is a color space of all human-visible colors defined by the International Commission on Illumination using three coordinates: L^* (lightness, 0–100), a^* (red–green axis), and b^* (yellow–blue axis). Following Wilson, Ghosh, and Caliskan (2025), we combine the a^* and b^* values for a comprehensive hue angle h° where $h^\circ = \arctan(b^*/a^*)$ around a color circle where 0° is red, 90° is yellow, 180° is green, and 270° is blue. C^* represents the saturation of a color calculated using the distance from the L^* axis: $C^* = \sqrt{a^{*2} + b^{*2}}$.

While analyses accounting for skin undertone biases represent important recent developments (Branigan et al. 2024), such distinctions may not yet be salient to policy-makers. Thus, we focus on hue's significance in representing skin tone homogeneity or diversity in model outputs. To measure skin tone diversity among groups, we use the hue angle H^* along with chroma (C^*) to calculate ΔE_{2000}, which measures the perceptible difference between two colors. Lower values are more similar and higher values are more perceptibly different. Schuessler (2025) provides a scale for general interpretation: values below one are not perceptible by the human eye, values below 50 are considered more similar than opposite, and a value of 100 indicates colors are exact opposites (red–green or blue–yellow in this space).

Skin Tone Results. Figure 5 shows a color space representation of each image generated in each proxy group. Across models, most skin tones with an L^* value below 40 are classified as Black. Although Stable Diffusion-3.5-Large-Turbo diverges the most from U.S demographic patterns (average TVD of 0.452), it produces images with higher skin tone diversity and with darker skin tones on average compared to the other tested models. Imagen-4 Fast is the most homogeneous in its skin tone depiction and its images are lighter on

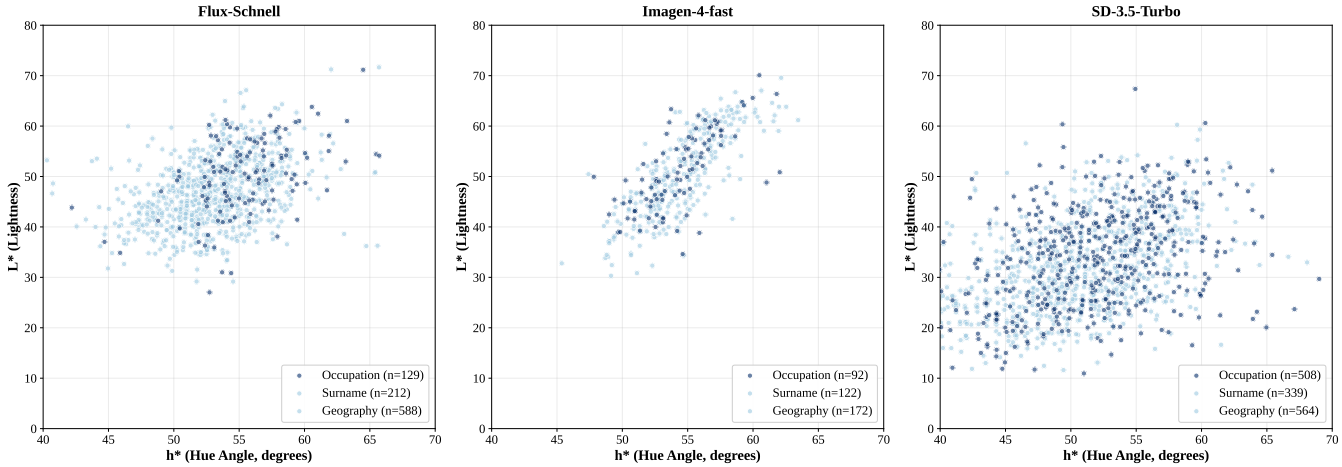


Figure 6: Black race only skin tone distributions across models and proxies

average than Flux Schnell and Stable Diffusion-3.5-Large-Turbo. This could partly be because it has the lowest number (172) of Black-classified images across the models tested. Though, Hispanic and Asian groups also have higher L^* values in Imagen-4 Fast.

Supplemental Figures S2 and S3 display the average L^* value and $\Delta E2000$ values across proxy category and race classification respectively. The supports the same conclusion as we found in Figure 5—for images classified as Hispanic and Asian, Imagen-4 Fast depicts lighter skin tones than Flux Schnell and Stable Diffusion-3.5-Large-Turbo; for images classified as Black, Stable Diffusion-3.5-Large-Turbo has an average L^* value of 31.5 compared to averages of 48.0 and 50.1 for Flux Schnell and Imagen-4 Fast. Stable Diffusion-3.5-Large-Turbo also shows higher skin tone diversity within races indicated by its higher on average ΔE values.

Figure 6 highlights trends for the measured skin tone for images classified as Black across the proxies. There is a trend, most noticeable in Flux Schnell, of occupational prompts producing images with higher L^* values compared to surname and geographic prompts. This is further examined in Figure 7 which displays each Black-classified image generated with occupational prompts. The specific prompts depicted are labeled at their average L^* value point against the average salary for the occupation (U.S. Census Bureau 2021a). The relationship between higher L^* values (lighter tones) with higher average salary is statistically significant, but weak. The skin tones of lower paying jobs are scattered, notably showing higher L^* values for stereotypically female occupations such as nail technician, nursing assistant, and elementary school teacher.

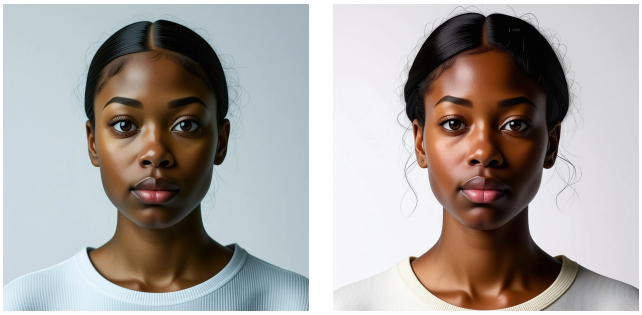


Figure 7: Correlation between salary and skin tone in Flux Schnell 's Black-classified outputs for occupation prompts.

Image Homogeneity within Groups

Diversity of output when using the same prompt is an objective measure of image quality in Text-to-Image design (OpenAI 2023). We measure image homogeneity to quantify if there is a difference in image quality within groups of images and also measure stereotypical representations in images. We draw upon the work of AlDahoul, Rahwan, and Zaki (2025) to measure visual homogeneity across generated images. We compute pairwise similarity within groups defined by race, prompt, and skin tone. Cosine similarity measures the angle between images embedding vectors, where values closer to 1.0 indicate that images are pointing in the same direction in embedding space and may be more visually alike. Two images with a high cosine similarity are shown in Figure 8. Higher intra-group similarity signals that a model is producing less diverse outputs for a given group. This lack of diversity within a group can be interpreted as a signal for stereotypical representation.

We extract two complementary embedding types to capture distinct elements of a model's output. We use CLIP embeddings (ViT-H-14, pretrained on LAION2B) that encode the overall appearance, style, and composition of a



(a) ”A person from Oakland” (b) ”A person from Queens”

Figure 8: Two images generated by Stable Diffusion-3.5-Large-Turbo with a cosine similarity of 0.838 between their facial embeddings.

photo alongside InsightFace buffalo_l face embeddings (InsightFace 2023) that encode facial morphology. This allows us to distinguish between homogeneity in the overall composition of an image versus the homogeneity in the faces and their features themselves. We filtered the dataset for face embeddings to those with a facial detection score above 0.85.

We define the *homogeneity gap* as the difference between mean intra-group and inter-group cosine similarity. A positive gap indicates that images within a group is more similar to each other than they are to images in other groups. While racial groups may cluster in embedding space due to shared physical features, our framework measures relative homogeneity among racial groups, prompts, and skin tones to find disproportionate self-similarity within groups.

Image Homogeneity Results. As shown in Figure 9, Black-classified images have higher intra-group similarity compared to other racial categories across all three models. This effect is most pronounced in Flux Schnell, which otherwise produces the most diverse facial embeddings but has a 0.110 facial homogeneity gap for Black-classified images. Asian-classified images show higher homogeneity gaps for CLIP embeddings (0.088 in Stable Diffusion-3.5-Large-Turbo, 0.106 in Imagen-4 Fast) than for facial embeddings (0.065 and 0.083 respectively), suggesting stylistic rather than facial homogeneity. Stable Diffusion-3.5-Large-Turbo notably produces more homogenous faces for White-classified images (gap: 0.107) than for Asian-classified images (0.065). In contrast to its diverse skin tone representation, Stable Diffusion-3.5-Large-Turbo has the highest CLIP and facial cosine similarity across models, with the highest average CLIP score (0.614 compared to 0.592 for Imagen-4 Fast and 0.500 for Flux Schnell).

Figure 10 shows the mean homogeneity gap for each prompt for Flux Schnell, organized by proxy category, demonstrating how our method enables analysis of representational differences across skin tone, prompt type, and specific prompt content. Like Flux Schnell, Imagen-4 Fast shows higher homogeneity for Asian-associated surname proxies, particularly Nguyen (0.239), Kim (0.237), and Tran (0.219). Flux Schnell produces higher homogeneity gaps for Black-associated geography proxies, with Memphis show-

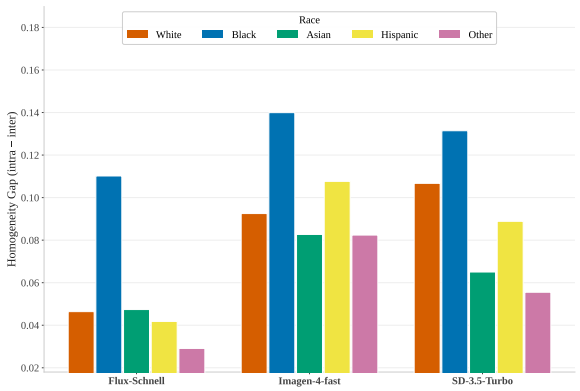


Figure 9: Homogeneity gap across models

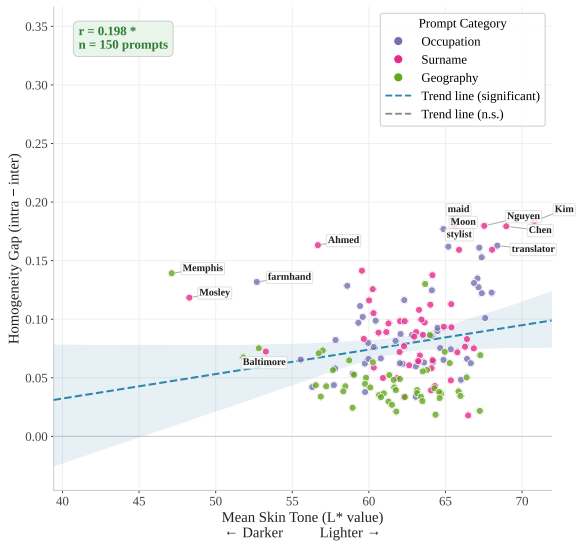


Figure 10: Facial embedding homogeneity gap by skin tone in Flux Schnell

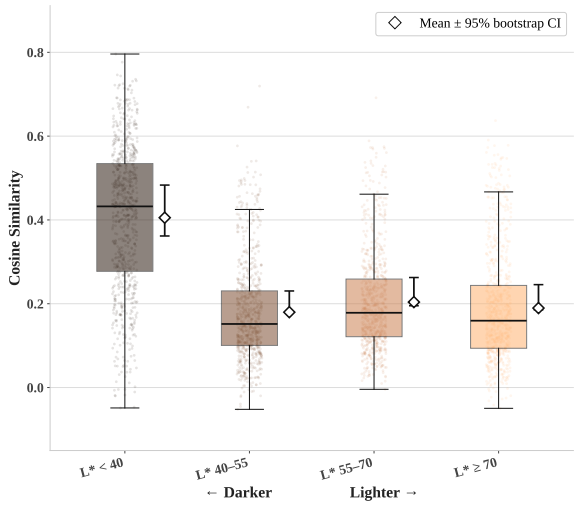


Figure 11: Imagen 4 fast facial embeddings cosine similarity by skin tone.

ing a gap of 0.140 and the darkest mean skin tone among top geography prompts ($L^* = 47.1$). These results demonstrate how proxies can uncover biases in model outputs that would not be indicated by racial classification metrics.

Analyzing images by skin tone bin rather than prompt-level averages reveals that darker-skinned images have higher intra-group facial similarity across all models and prompt categories. This effect is most pronounced in Imagen-4 Fast, as shown in Figure 11. Darker images with $L^* < 40$ achieve a mean facial similarity of 0.405, compared to 0.183 for $L^* \in [40, 55)$, a difference of 0.222. Flux Schnell shows a similar decline from 0.342 ($L^* < 40$) to 0.251 ($L^* \geq 70$). These results were found with a small sample size of 40 as Imagen 4 only generated 40 images with a skin tone with $L^* < 40$.

Discussion

Although bias in machine learning has been widely studied, integrating methods for measuring and mitigating biases into practical use in a regulated space remains challenging. Our proxy-based auditing method is based on established US anti-discrimination law and provides fine-grained data that can uncover fine-grained biases in model output. We use U.S. demographics, but our approach can be applied to any referenced demographic data distribution. Our method also reveals the extent a model's outputs reflect various demographic trends, providing interesting findings about image models that would not be revealed by traditional auditing methods.

We do not advocate for demographic accuracy as the preferred benchmark, but emphasize that our results demonstrate how models can amplify current disparities, and how this varies across proxies. Even demographically accurate outputs, however, can reinforce stereotypes and proliferate historical exclusion. For example, just five percent of U.S. lawyers are Black in 2024 (American Bar Association 2024). Should models reflect this current true representation or should they present an aspirational future with more equal representation?

These are not rhetorical questions. They impact procurement and design decisions that schools, institutions, and companies are making unwittingly or intentionally. School districts may prioritize local demographic accuracy; others may pursue aspirational representation especially for underrepresented groups. Auditing does not prescribe a particular answer, but it equips decision-makers with data to inform decisions and highlights issues that may not be apparent without effective auditing tools.

Limitations

Our work is limited by its U.S. focus. The proxies we employ are calibrated to American demographic data and legal frameworks. Adapting our method to other contexts would require using locally relevant proxies and demographic benchmarks, and many localities may not have appropriate granular demographic data available. Even U.S reference data is limited in its defined categories and its date of collection.

Racial classification presents inherent limitations as a measurement framework. Not only is race a subjective classification with no possible ground truth especially for artificially generated images, but the nature of racial classification leaves it as an unreliable measure for many demographic groups in America. Our methodology relies on the performance of a model for classification, and that model is limited by its own documented biases. For the model we used, results were especially unreliable for Hispanic-classified images. This is an imperfect attempt and automating the grouping of subjective classes. The model's tendency to classify more images as Hispanic than both human classifiers limits the accuracy of the results and TVD calculations. There are similar inherent limitations with our automated skin tone measurement which is subject to confounds based on lighting, irregular display, or shadows on faces that can lead to inaccuracies in measurement of skin tone.

Any results from our auditing system cannot be used without careful interpretation. While it is possible to use our system solely to examine TVD and analyze demographic accuracy, the true benefit comes from examining the data and analyzing the distribution and biases that exist in different models. There may be ways to simplify this analysis through a combined score or other measures to make the process easier to implement, but this would come at the expense of arguably necessary nuance.

Acknowledgments

This work is supported in part by funds provided by the National Science Foundation, Department of Homeland Security, and IBM through the ACTION AI Institute (Award #2229876). The authors thank Thomas Nachbar, who was instrumental in workshopping this project's initial ideas, and who offered valuable comments and suggestions throughout its development.

References

AlDahoul, N.; Rahwan, T.; and Zaki, Y. 2025. AI-generated faces influence gender stereotypes and racial homogenization. *Scientific Reports*, 15(14449).

Ameenuddin, N. 2025. Why #Representation Matters in Media. In *Untangling the Thread of Racism: A Primer for Pediatric Health Professionals*. American Academy of Pediatrics.

American Bar Association. 2024. Profile of the Legal Profession 2024.

Arif, M.; and Takefuji, Y. 2025. Why AI image generators cannot afford to be blind to racial bias. *AI & Society*, 40.

Baker, R.; and Hawn, A. 2022. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, 32: 1052–1092.

Barocas, S.; and Selbst, A. 2016. Big Data's Disparate Impact. *California Law Review*, 104: 671.

Bent, J. R. 2019. Is Algorithmic Affirmative Action Legal? Stetson University College of Law Research Paper No. 2019-6.

Bernard, D. L.; Smith, Q.; and Lanier, P. 2022. Racial discrimination and other adverse childhood experiences as risk factors for internalizing mental health concerns among Black youth. *Journal of Traumatic Stress*, 35(2): 473–483.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *ACM Conference on Fairness, Accountability, and Transparency*.

Branigan, A. R.; Nunez, J. G.; Khan, M. A.; and Gordon, R. A. 2024. Variation in Skin Red and Yellow Undertone: Reliability of Ratings and Relevance for Perceived Social Experiences. *Social Psychology Quarterly*, 87(3): 249–271.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability and Transparency*.

Castleman, J.; and Korolova, A. 2025. Adultification Bias in LLMs and Text-to-Image Models. In *ACM Conference on Fairness, Accountability, and Transparency*.

Cave, S.; and Dihal, K. 2020. The Whiteness of AI. *Philosophy & Technology*, 33: 685–703.

Cheong, M.; Abedin, E.; Ferreira, M.; Reimann, R.; Chalson, S.; Robinson, P.; Byrne, J.; Ruppanner, L.; Alfano, M.; and Klein, C. 2024. Investigating Gender and Racial Biases in DALL-E Mini Images. *ACM Journal on Responsible Computing*, 1(2): 13.

City of New York. 2021. New York City Administrative Code, Section 20-871: Requirements for Automated Employment Decision Tools.

City of New York. 2023. Rules of the City of New York, Title 6, Section 5-301: Bias Audit.

Clay-Gilmore, M. 2025. Visualizing the Enemy: Artificial Intelligence Image Generators and the Reproduction of Counter-Insurgent Logics in the Western Imagination.

Fitzpatrick, T. B. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology*, 124: 869–871.

Friedrich, F.; Brack, M.; Struppek, L.; Hintersdorf, D.; Schramowski, P.; Luccioni, S.; and Kersting, K. 2025. FairDiffusion: Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*.

Ghosh, S.; and Caliskan, A. 2023. 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. In *Findings of the Association for Computational Linguistics: EMNLP*.

Ghosh, S.; and Wilson, K. 2025. Bias Is a Math Problem, AI Bias Is a Technical Problem: 10-Year Literature Review of AI/LLM Bias Research Reveals Narrow [Gender-Centric] Conceptions of 'Bias', and Academia-Industry Gap. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Google AI & Ellis P. Monk. 2025. Monk Skin Tone Scale. https://skintone.google/.

Google Developers Blog. 2025. Announcing Imagen 4 Fast and the General Availability of the Imagen 4 Family in the Gemini API. https://developers.googleblog.com/announcing-imagen-4-fast-and-imagen-4-family-generally-available-in-the-gemini-api/.

Harris, C. L. 2025. Race, Representation, and Black Racial Identity Formation. *Professing Education*, 23(2).

InsightFace. 2023. buffalo_l: Facial Detection and Recognition Model. https://huggingface.co/immich-app/buffalo_l.

Karpouzis, K. 2024. Plato's Shadows in the Digital Cave: Controlling Cultural Bias in Generative AI. *Electronics*, 13(8): 1457.

Kim, P. T. 2017. Data-Driven Discrimination at Work. *William & Mary Law Review*, 58: 857.

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Lawrence, M.; Cimermanis, K.; and Collin, C. 2025. Not all AI-generated faces are created equal: impacts of model gender, race, and emotional expression on classification accuracy. *AI & Society*.

Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable Bias: Evaluating Societal Representations in Diffusion Models. In *Advances in Neural Information Processing Systems*.

Mbatha, S. K.; Booysen, M. J.; and Theart, R. P. 2024. Skin Tone Estimation under Diverse Lighting Conditions. *Journal of Imaging*, 10(5): 109.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6): 1–35.

Monk, E. P. 2023. The Monk Skin Tone Scale (MST). https://osf.io/preprints/socarxiv/pdf4c_v1.

NaviGator AI/Black Forest Labs. 2026. Flux.1 Schnell. https://docs.ai.it.ufl.edu/docs/navigator_models/models/blackforest-flux.1-schnell/.

NetChoice, LLC v. Bonta. 2024. 113 F.4th 1101, 1114 (9th Cir. 2024).

Nicoletti, L.; and Bass, D. 2023. Generative AI Takes Stereotypes and Bias From Bad to Worse. Bloomberg Graphics.

Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

OpenAI. 2023. DALL·E 3 System Card. https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf.

Park, Y. S. 2024. White Default: Examining Racialized Biases Behind AI-Generated Images. *Art Education*, 77(4): 36–45.

Pauker, K.; Williams, A.; and Steele, J. R. 2016. Children's Racial Categorization in Context. *Child Development Perspectives*, 10(1): 33–38.

Peinl, R. 2025. Using LLMs as prompt modifier to avoid biases in AI image generators. In *9th International Conference on Advances in Artificial Intelligence*.

Quintana, S. M. 1998. Children's Developmental Understanding of Race and Ethnicity. *Applied & Preventive Psychology*, 7(1): 27–45.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *ACM Conference on Fairness, Accountability, and Transparency*.

Schuessler, Z. 2025. Delta E 101 – Learn about Delta E Color Difference. https://zschuessler.github.io/DeltaE/learn/.

Shaw v. Reno. 1993. 509 U.S. 630, 631, 113 S. Ct. 2816, 2818, 125 L. Ed. 2d 511.

Siegel, R. B. 2004. Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over Brown. *Harvard Law Review*, 117(5): 1470–1547.

Sim, M.; Brigham, N. G.; Kohno, T.; Charlesworth, T. E. S.; and Caliskan, A. 2025. Biased AI Outputs Can Impact Humans' Implicit Bias: A Case Study of the Impact of Gender-Biased Text-to-Image Generators. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Simms v. First Gibraltar Bank. 1996. 83 F.3d 1546, 1551 n.12 (5th Cir. 1996).

Smedley, A.; and Smedley, B. D. 2012. *Race in North America: Origin and Evolution of a Worldview*. Boulder, CO: Westview Press. ISBN 9780813345543.

Spears Brown, C.; and Bigler, R. 2005. Children's Perceptions of Discrimination: A Developmental Model. *Child Development*, 76: 533–553.

Stability AI. 2024. Introducing Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5.

Steele, C. M.; and Aronson, J. 1995. Stereotype Threat and the Intellectual Test Performance of African Americans. *Journal of Personality and Social Psychology*, 69(5): 797–811.

Stuart v. Loomis. 2014. 992 F. Supp. 2d 585, 588 (M.D.N.C.), aff'd sub nom. Stuart v. Camnitz, 774 F.3d 238 (4th Cir. 2014).

Students Engaged in Advancing Texas v. Paxton. 2025. No. 1:24-CV-1326-DAE, 2025 WL 1007391, at *18–19 (W.D. Tex. Jan. 23, 2025).

Students for Fair Admissions, Inc. v. President and Fellows of Harvard College. 2023. Supreme Court of the United States.

U.S. Census Bureau. 2021a. Detailed Occupation by Education, Sex, Age, Race, and Earnings: ACS 2022.

U.S. Census Bureau. 2021b. Frequently Occurring Surnames from the 2010 Census.

U.S. Census Bureau. 2024a. About the Topic of Race. https://www.census.gov/topics/population/race/about.html.

U.S. Census Bureau. 2024b. DP05: ACS Demographic and Housing Estimates. 2024 ACS 1-Year Estimates Data Profiles.

U.S. Congress. 1964. Civil Rights Act of 1964, Title VII. 42 U.S.C. § 2000e et seq. Pub. L. No. 88-352, 78 Stat. 241.

U.S. Equal Employment Opportunity Commission. 2016. EEOC Enforcement Guidance on National Origin Discrimination.

Verkruijsse, W.; Brancart, A.; Jaffe, M. B.; and Groen, S. 2024. Variability of Printed Monk Skin Tone Scales May Cause Misclassification of Clinical Study Participants: Caveats On Printing. *Anesthesia & Analgesia*, 138(6): e43–e44.

Vázquez, A. F. d. C.; and Garrido-Merchán, E. C. 2024. A Taxonomy of the Biases of the Images Created by Generative Artificial Intelligence. *arXiv preprint arXiv:2407.01556*.

Wan, Y.; and Chang, K.-W. 2025. The Male CEO and the Female Assistant: Evaluation and Mitigation of Gender Biases in Text-To-Image Generation of Dual Subjects. In *63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Wang, J.; Liu, X.; Di, Z.; Liu, Y.; and Wang, X. 2023. T2IAT: Measuring Valence and Stereotypical Biases in Text-to-Image Generation. In *Findings of the Association for Computational Linguistics*.

Waters, M. 2024. Fair Admissions, Fair Decisions, and Fair Outcomes: An Analysis of Algorithmic Bias in Education, Employment, Healthcare, and Housing. *North Carolina Journal of Law & Technology*, 25.

Wilson, K.; Ghosh, S.; and Caliskan, A. 2025. Bias Amplification in Stable Diffusion's Representation of Stigma Through Skin Tones and Their Homogeneity. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Yang, Y. 2025. Racial Bias in AI-Generated Images. *AI & Society*, 40: 5425–5437.

Zhang, C.; Chen, X.; Chai, S.; Wu, C. H.; Lagun, D.; Beeler, T.; and la Torre, F. D. 2023. ITI-GEN: Inclusive Text-to-Image Generation. In *International Conference on Computer Vision*.

Zhou, M.; Abhishek, V.; Derdenger, T.; Kim, J.; and Srinivasan, K. 2024. Bias in Generative AI. arXiv:2403.02726.

Generative AI Disclosure Statement

Generative AI tools were used to assist with table formatting, code for data visualization, bibliography formatting, and light grammatical assistance. All of the text was written by the human authors with no AI-generated text, and all of the generated code was carefully reviewed for correctness.