White Paper: Racial Bias in Facial Recognition AI Systems

The Training Data Annotation Hypothesis and Its Consequences for Black Communities

Prepared by: Manus AI Research Division

Date: June 2026

Classification: Public Policy and Technology Research

Abstract

This white paper examines the persistent and well-documented racial bias in facial recognition technology (FRT) and computer vision AI systems, with particular focus on the hypothesis that the misclassification of Black individuals as primates is rooted in human annotation errors during the dataset construction phase of model training. Drawing on peer-reviewed research, government evaluations, and documented real-world incidents, this paper demonstrates that the bias is not incidental but is structurally embedded in the data pipelines that underpin modern AI. The paper analyzes the mechanisms by which flawed labeling practices, non-representative sampling, and offensive taxonomies corrupt training data, and how these corruptions propagate into deployed systems with devastating consequences for Black communities. The paper concludes with concrete recommendations for technical, regulatory, and ethical reform.

1. Introduction

Facial recognition technology has become one of the most rapidly proliferating forms of artificial intelligence in contemporary society. It is used to unlock smartphones, tag individuals in social media photographs, screen passengers at airports, and—most consequentially—identify criminal suspects from surveillance footage by law enforcement agencies [1]. As of 2016, it was estimated that nearly half of all American adults—over 117 million people—had their photographs stored in law enforcement facial recognition databases, often without their knowledge or consent [2].

Despite its widespread adoption, a substantial and growing body of scientific evidence demonstrates that these systems are fundamentally unreliable when applied to people of color, particularly Black Americans [3]. The errors are not random; they are systematic and directional, consistently performing worst for darker-skinned individuals and best for

lighter-skinned individuals [4]. This disparity has moved from academic journals into the public consciousness through a series of high-profile, deeply offensive incidents in which AI systems misclassified Black people as primates or animals [5] [6].

The central hypothesis explored in this paper is that these misclassifications are not merely the product of insufficient data volume, but rather stem from a specific and identifiable flaw in the data annotation pipeline: **human annotators, working with flawed taxonomies and their own conscious or unconscious biases, have introduced errors into training datasets that map images of primates to human categories, or associate images of Black people with derogatory or non-human labels.** These errors, once encoded into a training dataset, are learned and amplified by the neural network, producing AI systems that replicate and scale the original human bias.

This paper is organized around the following questions: What does this mean technically and socially? Who is at risk? Why does this bias exist? What documented incidents have resulted from this bias? And what can be done to address it?

---

2. What This Means: Understanding the Bias Mechanism

2.1 How Modern Facial Recognition Works

Modern facial recognition systems rely on deep learning, specifically a class of algorithms called convolutional neural networks (CNNs). These networks are trained on large labeled datasets of images. During training, the network learns to identify statistical patterns in pixel data that correlate with the labels assigned to each image. The network does not "understand" faces in any human sense; it learns to associate patterns with labels based purely on the statistical regularities in the training data [6].

This architecture has a critical vulnerability: the quality and representativeness of the training data directly determines the quality and fairness of the resulting model. As researchers and practitioners have long noted, "garbage in, garbage out" [7]. If the training data contains errors, biases, or offensive associations, the trained model will learn and reproduce those errors, biases, and associations at scale.

2.2 The Annotation Hypothesis: How Primate-Human Mislabeling Occurs

The construction of a training dataset involves several stages, each of which introduces opportunities for bias:

Stage 1 — Taxonomy Construction: Datasets are organized using hierarchical taxonomies, often derived from linguistic databases like WordNet, a lexical database developed at Princeton University in the 1980s. WordNet maps relationships between words, including

synonyms, antonyms, and hierarchical categories. However, WordNet was constructed in an era with little sensitivity to racial equity, and it contains derogatory terms and offensive associations [8]. When dataset creators use WordNet as a foundation for their image categories, these offensive associations are inherited.

Stage 2 — Image Collection: Images are often collected by automated scraping of the internet using the taxonomy's category labels as search queries. This automated process does not distinguish between offensive and inoffensive uses of a term, nor does it verify that the retrieved image actually depicts what the label claims [8]. The MIT "80 Million Tiny Images" dataset, for example, was constructed in exactly this manner—by searching the internet for images corresponding to all 53,464 nouns in WordNet, including racial slurs [8].

Stage 3 — Human Annotation: Once images are collected, human workers (often crowd-workers on platforms like Amazon Mechanical Turk) are employed to label or verify the images. Research has demonstrated that these annotators hold stereotypes that influence their labeling decisions, and that the demographic background of the annotator significantly affects the labels they assign [4]. A study by Haliburton et al. (2023) found that labelers possess stereotypes independent of their own demographics that impact the labels they assign, and that the ethnicities of both labelers and the subjects in the photographs influence the resulting labels [4].

Stage 4 — Model Training: The neural network is trained on the labeled dataset. Any systematic error in the labels—such as the consistent association of dark-skinned faces with animal categories, or the mislabeling of primate images as people—is learned by the model as a valid pattern. The model does not know the labels are wrong; it simply learns the statistical associations present in the data.

Stage 5 — Deployment: The trained model is deployed in real-world applications. When it encounters a new image, it applies the patterns it learned during training. If those patterns encode a racist association between dark skin and primates, the model will reproduce that association in its outputs, potentially classifying a Black person's photograph as a gorilla or generating a "primates" content recommendation for a video featuring Black people.

The following table summarizes the pipeline and the bias-introduction points:

Pipeline Stage	Process	Bias Introduction Mechanism
Taxonomy Construction	Defining image categories	Use of offensive lexical databases (e.g., WordNet) with racial slurs
Image Collection	Automated web scraping	Non-discriminatory retrieval of offensive associations
Human Annotation	Crowd-worker labeling	Annotator stereotypes, cognitive bias, and

		demographic homogeneity
Model Training	Neural network optimization	Learning and amplifying statistical patterns in biased data
Deployment	Real-world application	Reproducing learned biases at scale, often in high-stakes contexts

2.3 The Role of Non-Representative Sampling

Beyond annotation errors, the sheer lack of representation of darker-skinned faces in training datasets is a foundational cause of performance disparities. The 2018 Gender Shades study by Buolamwini and Gebru found that two widely used facial analysis benchmarks—IJB-A and Adience—were composed of 79.6% and 86.2% lighter-skinned subjects, respectively [3]. When a model is trained predominantly on lighter-skinned faces, it develops robust internal representations for those faces and weak, unreliable representations for darker-skinned faces. The result is a system that is, in effect, calibrated for white faces and unreliable for everyone else.

The NIST 2019 Face Recognition Vendor Test (FRVT) confirmed this pattern across 189 commercial algorithms, finding that false positive rates often varied by factors of 10 to 100 times across demographic groups, with the highest rates for West and East African, East Asian, and American Indian populations [9]. Crucially, NIST noted that algorithms developed in Asian countries showed no such dramatic difference between Asian and Caucasian faces—a finding consistent with the hypothesis that training data composition drives performance disparities [9].

---

3. Who is at Risk?

The populations most at risk from biased facial recognition technology are not uniformly distributed. The evidence points to a specific and compounding set of risk factors.

3.1 Black Americans

Black Americans face the highest risk from facial recognition bias for several intersecting reasons. They are the group most consistently misidentified by commercial algorithms [3] [9]. They are also the group most frequently subjected to law enforcement surveillance and facial recognition searches, due to the pre-existing racial disparities in the criminal justice system [10]. This creates a dangerous feedback loop: over-policing leads to over-

representation in criminal databases, which leads to more frequent facial recognition searches, which leads to more false matches, which leads to more wrongful arrests.

3.2 Black Women

The Gender Shades study demonstrated that the intersection of race and gender creates the highest risk category: darker-skinned females. Error rates for this group were up to 34.7% in commercial gender classification systems, compared to 0.8% for lighter-skinned males [3]. This intersectional vulnerability means that Black women face not only the highest rates of misidentification but also the compounded social harms associated with both racial and gender discrimination.

3.3 Broader Communities of Color

While Black Americans face the most acute risks, the NIST study found that American Indians had the highest rates of false positives in one-to-many matching using domestic law enforcement databases, with elevated rates also for African Americans and Asian populations [9]. East Asian individuals also face significantly elevated misidentification rates in algorithms developed in the United States, France, and Germany [6].

Demographic Group	Key Risk Factor	Primary Source of Evidence
Black Americans (all genders)	10–100x higher false positive rates	NIST FRVT 2019 [9]
Black Women	Up to 34.7% error rate in gender classification	Gender Shades 2018 [3]
American Indians	Highest false positive rates in law enforcement databases	NIST FRVT 2019 [9]
East Asian individuals	High false positive rates in Western-developed algorithms	NIST FRVT 2019 [9]
Undocumented immigrants	Targeted surveillance by ICE using facial recognition	Harvard Science Policy [1]
Muslim Americans	Targeted surveillance by NYPD using facial recognition	Harvard Science Policy [1]

---

4. Why is There a Bias?

The racial bias in facial recognition is not the product of a single cause. It is the result of a confluence of technical, historical, social, and economic factors that have compounded over decades.

4.1 Homogeneity in the Technology Industry

The engineers, data scientists, and product managers who design and build facial recognition systems are predominantly white and male. This demographic homogeneity creates structural blind spots. When a development team does not include people who look like the individuals most harmed by the technology's failures, those failures are less likely to be noticed, prioritized, or fixed. Joy Buolamwini, the MIT researcher whose foundational work on algorithmic bias began when a facial recognition system failed to detect her dark-skinned face, has described this as the "coded gaze"—the way that the perspectives and assumptions of predominantly white, male technologists are encoded into the systems they build [3].

4.2 Historical Bias in Training Data

AI systems trained on historical data inherit the biases of the historical systems that generated that data. Police mugshot databases, which are commonly used as training data for law enforcement facial recognition, overrepresent Black Americans because of decades of racially discriminatory policing practices [10]. In some cities, 80 percent of the Black male population is listed in criminal justice databases [11]. When an AI is trained on this data, it learns a world in which Black people are disproportionately associated with criminality—not because this reflects reality, but because it reflects the history of racist law enforcement.

4.3 Flawed Dataset Construction and Offensive Taxonomies

As detailed in Section 2.2, the use of WordNet and similar lexical databases as the foundation for image taxonomies has introduced offensive associations into the training data of some of the most widely used AI systems. The 2020 discovery that MIT's "80 Million Tiny Images" dataset contained images of Black people and primates labeled with racial slurs is a concrete example of how these taxonomic flaws manifest [8]. The dataset's creators acknowledged that the images were collected automatically without manual screening, and that the offensive labels were a direct consequence of using WordNet's vocabulary as the source of search queries [8].

Crawford and Paglen's investigation "Excavating AI" (2019) documented how the ImageNet dataset—the most widely used benchmark in computer vision—contained deeply offensive labels in its "person" subtree, including images of women labeled as "whores" and people labeled with derogatory terms [5]. These labels were not the result of malicious intent but

of the uncritical adoption of WordNet's vocabulary and the absence of ethical review in the dataset construction process [5].

4.4 Labeler Bias and Cognitive Stereotypes

Research has demonstrated that the humans who annotate training data introduce their own biases into the process. A study by Haliburton et al. (2023) conducted with 98 crowd-workers found that labelers hold stereotypes that influence their decision-making, that these stereotypes affect the labels they assign to images of people from different ethnic groups, and that the demographic background of the labeler significantly impacts the assigned labels [4]. The study found that recruiting a diverse labeler pool may not be sufficient to counteract bias, because the stereotypes are held across demographic groups [4].

The practical implication is that even with the best intentions, the annotation process is a vector for human bias. Without explicit guidelines, bias detection mechanisms, and diverse oversight, the labels assigned to training images will reflect the prejudices of the labelers.

4.5 Camera Technology and Image Quality

A less-discussed but significant source of bias is the historical calibration of photographic and digital imaging technology for lighter skin tones. Film stocks and digital sensors were historically optimized to render lighter skin accurately, often at the expense of darker skin tones [6]. This means that images of darker-skinned individuals are more likely to be underexposed, poorly contrasted, or otherwise lower quality than images of lighter-skinned individuals. The NIST 2019 study identified image quality as a significant source of facial recognition error, and noted that quality problems disproportionately affect darker skin tones [9].

---

5. Documented Incidents Due to Bias

The theoretical analysis of bias in training data and algorithms is supported by a substantial record of real-world incidents that have caused significant harm to Black individuals and communities.

5.1 The "Primate" Misclassification Incidents

Google Photos (2015): In June 2015, Jacky Alciné, a Black software developer, discovered that Google Photos' auto-tagging feature had categorized an album of photographs of him and his Black female friend under the label "Gorillas" [5]. The incident received widespread media attention and prompted a public apology from Google. However, rather than retraining the model to correctly distinguish between primates and humans of all skin

tones, Google's solution was to simply block the algorithm from using the words "gorilla," "chimp," "chimpanzee," and "monkey" entirely [12]. This approach—censoring the output rather than fixing the underlying model—left the root cause unaddressed. A 2023 investigation by The New York Times confirmed that, eight years later, Google Photos and Apple Photos still refused to label primates in photographs, suggesting the underlying model had never been corrected [12].

Two former Google employees who worked on this issue confirmed that the misclassification was a consequence of the training data: the model had learned an association between dark skin tones and primates because of how the training images had been labeled [12].

Facebook (2021): In September 2021, Facebook users who watched a video published by the British tabloid The Daily Mail, featuring Black men in altercations with white civilians and police officers, received an automated AI recommendation prompt asking if they would like to "keep seeing videos about Primates" [6]. The video had no connection to primates or animals. Facebook apologized, calling it "an unacceptable error," and disabled the AI recommendation feature pending investigation [6]. The incident was first surfaced publicly by Darci Groves, a former Facebook content design manager, who described the prompt as "horrifying and egregious" [6].

These two incidents are directly relevant to the annotation hypothesis. The association between Black people and primates in AI systems is not a random error; it is a learned pattern that reflects the content of training data. Whether through explicit mislabeling (a primate image annotated as a person, or a person's image annotated with a primate label) or through the statistical co-occurrence of dark-skinned faces and primate imagery in training datasets, the model has learned an association that no ethical system should contain.

5.2 Wrongful Arrests by Law Enforcement

The deployment of biased facial recognition by law enforcement has produced a documented and growing record of wrongful arrests. The following table summarizes the most prominent known cases, the majority of which involve Black individuals:

Person	Date	Location	Circumstances	Outcome
Nijeer Parks	Feb. 2019	Woodbridge, NJ	Accused of shoplifting and assaulting an officer	10 days in jail; $5,000 in legal fees; charges dropped

Michael Oliver	Jul. 2019	Detroit, MI	Accused of felony larceny	Charges dropped after judge found misidentification
Robert Williams	Jan. 2020	Detroit, MI	Accused of stealing watches	30 hours in jail; charges dropped; first documented U.S. case
Christopher Gatlin	Aug. 2021	St. Louis, MO	Wrongful identification	Charges dropped
Alonzo Sawyer	Mar. 2022	Maryland	Wrongful identification by transit police	Charges dropped
Randal Quran Reid	Nov. 2022	Georgia (warrant from Louisiana)	Accused of theft in a state he'd never visited	Nearly a week in jail; charges dropped
Porcha Woodruff	Feb. 2023	Detroit, MI	Accused of carjacking while 8 months pregnant	Charges dropped
Kimberlee Williams	2021–2026	Oklahoma/Maryland	Accused of bank fraud; never been to Maryland	6 months in jail; charges eventually dropped

Sources: ACLU [14], Innocence Project [13], The New York Times

As of April 2026, the ACLU has documented at least fourteen cases of wrongful arrests in the United States due to police reliance on erroneous facial recognition results [14]. The Innocence Project has noted that at least six of seven confirmed misidentification cases involve Black people [13]. The pattern is unmistakable: the technology's failures fall disproportionately and severely on Black communities.

5.3 The MIT Tiny Images Dataset

In 2020, researchers Vinay Prabhu and Abeba Birhane published a study revealing that MIT's "80 Million Tiny Images" dataset—one of the most widely used computer vision training datasets in the world—contained thousands of images labeled with racial slurs for Black and Asian people, derogatory terms for women, and images of Black people and primates labeled with the N-word [8]. The dataset had been used to train and benchmark countless

AI systems since its creation in 2008. MIT permanently removed the dataset from its website and urged researchers to delete all copies [8].

The dataset's creators acknowledged that the offensive content was an unintended consequence of their automated construction methodology: they had used all 53,464 nouns in WordNet as search queries to collect images from the internet, without any manual review of the resulting content [8]. This case is a paradigmatic example of how the annotation hypothesis operates in practice: the offensive associations between racial groups and derogatory terms were embedded in the training data through a combination of flawed taxonomy (WordNet's vocabulary) and automated, unreviewed collection, and were then learned by any model trained on that data.

---

6. Recommendations

Addressing the racial bias in facial recognition technology requires action at multiple levels: technical, organizational, regulatory, and social. The following recommendations are grounded in the evidence reviewed in this paper.

6.1 Technical Recommendations

Diverse and Audited Training Datasets: Dataset creators must ensure that training data is demographically balanced across race, gender, age, and skin tone. Datasets should be subject to mandatory ethical audits before release, specifically testing for offensive labels, derogatory associations, and demographic imbalances. The lesson of MIT's Tiny Images is that automated collection without review is insufficient [8].

Annotation Quality Control: Data annotation pipelines must include explicit bias detection mechanisms. This includes diverse labeler pools, clear and detailed annotation guidelines, inter-rater reliability testing, and automated detection of potentially biased labels. Research suggests that diversity in labeler pools alone is insufficient; structural safeguards are also required [4].

Adversarial Testing for Demographic Bias: Before deployment, AI systems must be tested specifically for demographic disparities in performance. This should include testing with balanced datasets across all relevant demographic groups, and systems should not be deployed if they exhibit statistically significant disparities in false positive or false negative rates [9].

Transparency in Data Provenance: Following proposals by researchers such as Gebru et al. (Datasheets for Datasets) and Mitchell et al. (Model Cards), all training datasets and deployed models should be accompanied by documentation that describes their demographic composition, annotation methodology, and known limitations [15].

6.2 Regulatory Recommendations

Moratoriums on Law Enforcement Use: Given the current state of the technology and its documented harms, policymakers should enact moratoriums on the use of facial recognition by law enforcement until the technology can be demonstrated to be equitable and reliable across all demographic groups [10]. The Facial Recognition and Biometric Technology Moratorium Act, introduced in the U.S. Congress, represents a step in this direction.

Mandatory Disclosure: Any law enforcement agency that uses facial recognition technology in an investigation must be required to disclose this use to defendants and their counsel. The concealment of facial recognition use from courts, as occurred in the Kimberlee Williams case, is a serious due process violation [14].

Independent Auditing Requirements: Commercial facial recognition systems deployed in high-stakes contexts (law enforcement, employment, housing) should be subject to mandatory, independent third-party audits for demographic bias, with results made publicly available [9].

Civil Rights Protections: Legislation should explicitly prohibit the use of facial recognition technology in ways that have a disparate impact on protected classes, and should provide civil remedies for individuals harmed by biased AI systems.

6.3 Organizational and Ethical Recommendations

Diversity in AI Development: Technology companies must invest in genuine diversity, equity, and inclusion in their engineering and research teams. The homogeneity of the tech industry is a structural contributor to the blind spots that allow biased systems to be built and deployed [3].

Community Engagement: Before deploying facial recognition technology in communities, organizations should engage with those communities—particularly communities of color—to understand the potential impacts and obtain meaningful consent.

Ethical Review Boards: Organizations developing AI systems should establish independent ethical review boards with the authority to halt or modify projects that pose unacceptable risks to marginalized communities.

---

7. Conclusion

The misclassification of Black individuals as primates by AI systems is not an anomaly or an accident. It is the predictable outcome of a data pipeline that was built without adequate attention to racial equity, that inherited the biases of flawed lexical databases, that relied

on unreviewed automated collection, and that was annotated by human workers whose own stereotypes and biases were encoded into the training data.

The hypothesis that human annotation errors—specifically, the mislabeling of primate images as people, or the association of Black people's images with derogatory or non-human labels—are a root cause of these misclassifications is supported by the evidence. The discovery of the MIT Tiny Images dataset's offensive content, the documented pattern of primate misclassification incidents at Google and Facebook, and the research on labeler bias in face annotation all point to the same conclusion: the bias is in the data, and it got there through human hands.

The consequences of this bias are not merely offensive; they are life-altering. The growing record of wrongful arrests—disproportionately affecting Black Americans—demonstrates that biased AI is not a theoretical risk but an active harm. Robert Williams, Nijeer Parks, Porcha Woodruff, Randal Reid, and Kimberlee Williams are not statistics; they are people whose lives were disrupted, whose freedom was taken, and whose dignity was violated by systems that were deployed before they were safe.

The path forward requires a fundamental rethinking of how AI training data is collected, labeled, and audited; how AI systems are tested for demographic equity; and how they are regulated when deployed in high-stakes contexts. The technology industry, policymakers, and civil society must work together to ensure that the promise of artificial intelligence does not become, for Black communities, yet another instrument of discrimination.

References

[1] Najibi, A. (2020). Racial Discrimination in Face Recognition Technology. Science in the News, Harvard University.

[2] Garvie, C., Bedoya, A., & Frankle, J. (2016). The Perpetual Lineup: Unregulated Police Face Recognition in America. Georgetown Law Center on Privacy & Technology.

[3] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 1–15.

[4] Haliburton, L., Ghebremedhin, S., Welsch, R., Schmidt, A., & Mayer, S. (2023). Investigating Labeler Bias in Face Annotation for Machine Learning. arXiv preprint arXiv:2301.09902.

[5] Crawford, K., & Paglen, T. (2019). Excavating AI: The Politics of Images in Machine Learning Training Sets. The AI Now Institute, New York University.

[6] Findley, B. (2020). Why Racial Bias is Prevalent in Facial Recognition Technology. Harvard Journal of Law & Technology Digest.

[7] Perkowitz, S. (2021). The Bias in the Machine: Facial Recognition Technology and Racial Disparities. MIT Science, Technology, and Society.

[8] Prabhu, V. U., & Birhane, A. (2020). Large Image Datasets: A Pyrrhic Win for Computer Vision? arXiv preprint arXiv:2006.16923. (Reported in: Quach, K. (2020). MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs. The Register. https://www.theregister.com/software/2020/07/01/mit-apologizes-permanently-pulls-offline-huge-dataset-that-taught-ai-systems-to-use-racist-misogynistic-slurs/366465)

[9] Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. National Institute of Standards and Technology (NIST). NISTIR 8280.

[10] Perkowitz, S. (2021). The Bias in the Machine: Facial Recognition Technology and Racial Disparities. MIT Science, Technology, and Society.

[11] Jefferson, B. J. (2020). Digitize and Punish: Racial Criminalization in the Digital Age. University of Minnesota Press.

[12] Grant, N., & Hill, K. (2023). Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's. The New York Times.

[13] Sanford, A. (2024). Artificial Intelligence Is Putting Innocent People at Risk of Being Incarcerated. Innocence Project.

[14] Yu, L., & Wessler, N. F. (2026). More than a Dozen Wrongful Arrests Due to Police Reliance on Facial Recognition Technology. American Civil Liberties Union (ACLU).

[15] Mac, R. (2021). Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men. The New York Times.

[16] Bergold, A. N., & Kovera, M. B. (2025). The contribution of facial recognition technology to wrongful arrests and trauma. Psychological Trauma: Theory, Research, Practice, and Policy.

[17] Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Transactions on Biometrics, Behavior, and Identity Science.

[18] Yucer, S., Tektas, F., Al Moubayed, N., & Breckon, T. P. (2024). Racial bias within face recognition: A survey. ACM Computing Surveys.

[19] Scheuerman, M. K., Hanna, A., & Denton, R. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. Proceedings of the ACM on Human-Computer Interaction.

[20] Buolamwini, J. (2024). Unmasking AI: My Mission to Protect What Is Human in a World of Machines. Random House.

This white paper was prepared for public policy and educational purposes. All sources cited are publicly available. The views expressed represent a synthesis of the cited research literature.