Cognitec White Paper

Bias in Face Recognition Systems: Controversial Opinions and Proven Facts

The list of justifications for recent bans of face recognition technologies always includes bias—the assumption that all systems produce higher error rates for certain demographic groups, mainly people of color, and therefore cause wrongful arrests, human rights issues and other unfair treatment.

However, reputable tests show that high-performing facial matching algorithms also exhibit extremely low accuracy differences when matching images from certain ethnicities and countries.

But slight differences do exist. Reasons for this behavior do not always lie in the technology itself, but often in image quality problems. And, different studies yield differing results, underlining the complexity of accuracy tests and the falsity of generalized bias statements.

A nuanced, scientific evaluation of currently available test data, as presented in this paper, serves to debunk the fundamental misrepresentations that face recognition algorithms are biased, and should demonstrate their technical capabilities to work with and for diverse populations.

Note: Accuracy evaluations for **face recognition** algorithms that perform **image matching** must not be confused with results from **gender and ethnicity classification** studies.

The National Institute of Standards and Technology (NIST) uses the term **demographic differences** (not bias) to describe performance variations on population groups.

NIST and MdTF Tests

The NIST Face Recognition Vendor Test (FRVT): Demographic Effects, published in December 2019, performed a multitude of tests with 18 million images of 8.5 million subjects to determine the influence of demographics (age, gender, ethnicity, country of birth) on matching accuracy.

The main finding of the tests, as summarized by the Biometrics Institute, indicates that women in general produce higher false non-match rates. However, this is a "marginal effect"—98% of women are still correctly verified. Conversely, fewer than 2% of comparisons fail to verify the person.

In addition, false positive rates are highest in West/East African and East Asian people and lowest in Eastern Europeans. When evaluating higher-quality photographs from a global population of visa applicants, false positives are also higher in women than in men and elevated for the elderly and children.

Aside from fluctuating algorithm performance, matching errors stem from varying degrees of image quality and lighting. With U.S. domestic mugshots taken with a photographic setup to produce high-quality images, false negatives are higher in Asian and American Indians. But using lower-quality U.S. border crossing images, false negatives are higher in people born in Africa or the Caribbean.

This implies that images of black people may yield higher inaccuracies not due to algorithm failure, but because poorly lit faces (not photographed with sufficient lighting, or with lighting that produces hot spots) show a low number of gray-scale values, and therefore an insufficient dynamic range to work well for matching algorithms.

The International Biometrics and Identity Association (IBIA) stated that the NIST test showed wide performance variations, ranging from algorithms that are "less accurate than a coin toss, to high performing algorithms that are overwhelmingly accurate with virtually undetectable demographic differences." And, these latter algorithms are 20 times more accurate than the most highly-skilled human adjudicators.

Papers published by the Identity and Data Sciences Laboratory at the Maryland Test Facility (MdTF) have explored the role of image acquisition on demographic differences and system performance, and the influence of demographics on false match rate (FMR) estimates for facial recognition systems.

In a test simulating an unattended, high-throughput scenario, MdTF found that many combinations of image acquisition systems and matchers met the 95 percent true identification rate (TIR) across all racial groups. Most of the errors were not made by algorithms, but at the image acquisition stage. For people wearing face masks, the various systems uniformly showed higher error rates for people of color, for both capturing and matching processes.

NIST test results for Cognitec technology

NIST used algorithms submitted to 1:1 and 1:N FRVTs for a specific study on demographic effects, and in 2019 evaluated their performance on the following image sets: US mugshots, immigration application images, visa application images, and border crossing images.

Cognitec's performance in the identification test, for example, using U.S. mugshots, shows very low influence of demographics on matching accuracy for white males vs. black males, and for white males vs. white females. However, black females produced more false matches than black males.

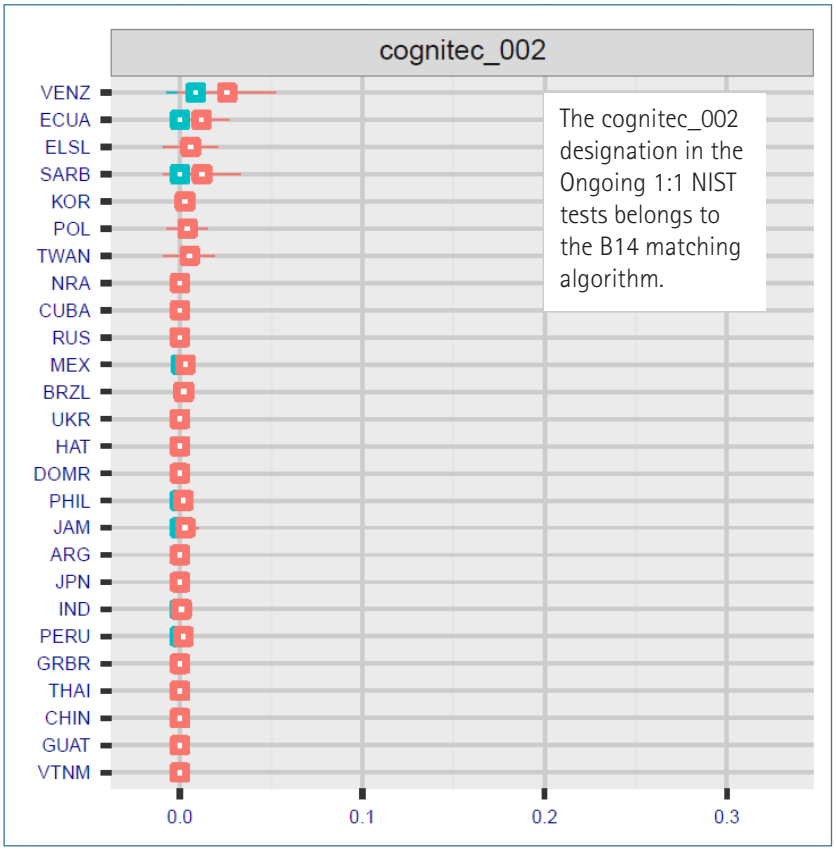
While demographics have an influence on the performance of all leading algorithms, and many show similar error rates for certain demographic groups, the absolute error rates are low.

For example, if the threshold is set for an overall FMR of 0.01% = 1 person in 10,000, the false match rate (FMR) for the most difficult demographic (women from West Africa) is 0.31% = 31 persons in 10,000 (see figure 9, page 43).

In real-world applications, users can set separate matching thresholds for certain demographic groups to eliminate the issues with FMR variations.

A subtest on visa images in the NIST 1:1 FRVT report from March 2021 shows the distribution of false non-match rates (FNMR) across several countries of birth, for two fixed false match rates (FMR). The effects are likely due to image quality variations, rather than demographics like age and race.

In the graph below (see figure 178, page 222), the blue point refers to a match threshold corresponding to a FMR=0.1%, and the red point to a FMR=0.01%. The FNMRs are uniformly low. The highest FNMRs for Cognitec's latest matching algorithm, B14, are observed for Venezuela, for unknown reasons, perhaps rooted in the quality of the visa photos. For this country of birth, the FNMR is about 0.88% at a FMR of 0.1%.



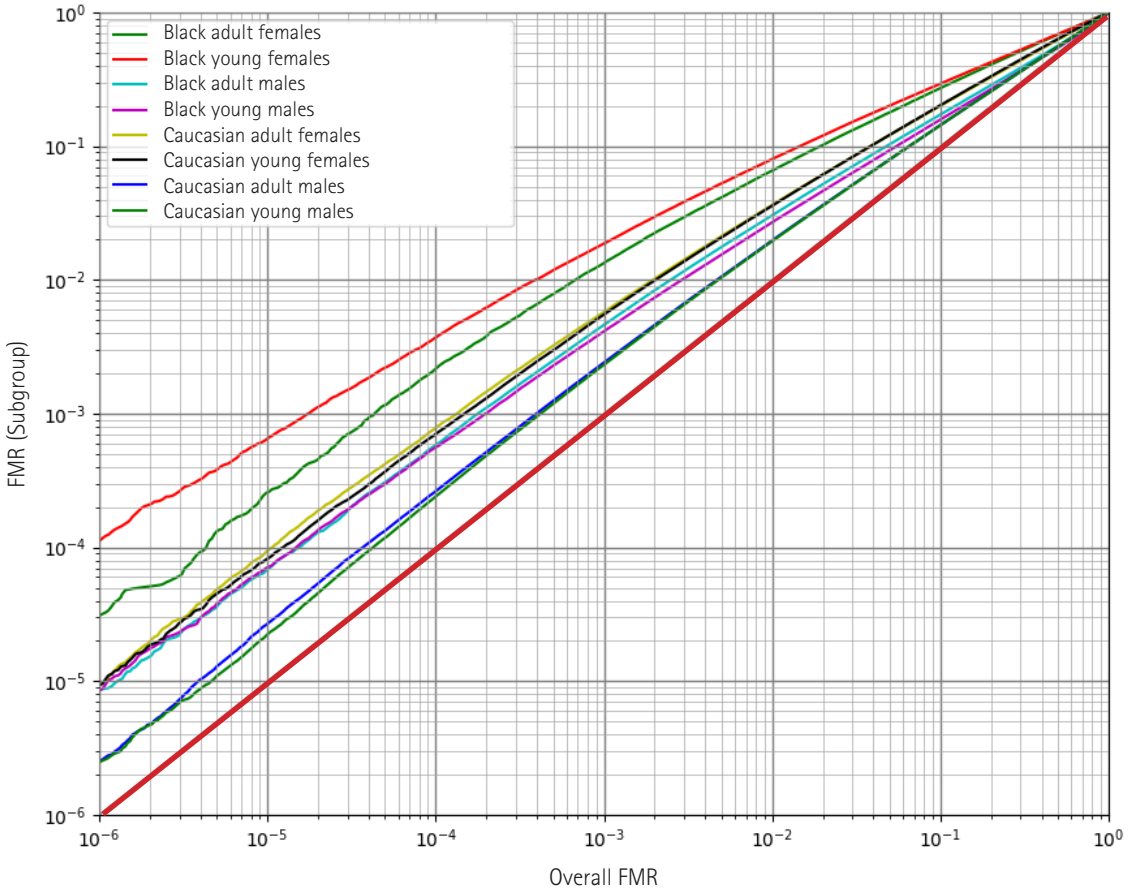
Cognitec in-house tests

In early 2021, Cognitec conducted in-house tests on demographic effects with the company's latest face matching algorithm B14.

The test suite measured the difference in false match rate and false non-match rate between the baseline (all samples of all subjects together) and subsets of imposter and genuine comparisons, respectively.

The subsets are chosen for comparisons within certain demographic groups, similar to the method used for the NIST report on demographic effects. Cognitec used an independent set of U.S. mugshots, and defined groups according to ethnicity, gender and age of the subjects.

The test results below show the FMR for all groups, as a function of the overall FMR for B14. The closer a group curve is to the diagonal (in red), the less the FMR is elevated for this demographic group.



In general, Cognitec's tests confirmed results from the NIST report: black females show relatively higher FMRs, when using the whole data with the same matching threshold. Overall error rates are reasonably low.

In addition, the in-house tests can give users of Cognitec's technology valuable guidance on how thresholds can be adapted for a variety of demographic groups. Users should involve face recognition experts during tests for demographic effects on their own image database and for optimal threshold configurations.

ISO standards and assessment metrics

The ISO 19795 series provides a framework for biometric system testing and evaluation, and Part 10 will apply to performance variations across demographic groups. The first draft is expected to be completed in 2021, with the final version anticipated to be published in 2023 or 2024.

Cognitec

Human bias

Racial biases are a human condition, often caused by lack of information. Diversely trained facial recognition algorithms use immense data to make neutral decisions, void of human prejudice. Cognitec's algorithms, for example, do not use any skin color information, and can therefore support more impartial investigations.

Georgia Tech College of Computing Ph.D. alumna Samira Samadi argued that "human-subject experiments should be looked into before concluding that human intervention is the silver bullet in solving software limitations."

Algorithms often behave similar to intuitive human behavior. For example, the probability of wrongly matching sets of people of the same sex, age and ethnicity is much higher than it is across a diverse population.

Face recognition bias in papers and media

Various papers and studies on bias in face recognition algorithms used poor accuracy results from tests classifying gender and skin type, instead of test results for matching performance.

Consequently, media discussions, political agendas and privacy associations rail terms like "techno-racism," and continuously cite two studies that have shown poor accuracy of gender classification algorithms on women with dark skin.

Those studies did not use mainstream, for-sale algorithms and applied very small sample sizes. The results showed that black women were male 35% of the time—a high number that led to quick, general and wrong conclusions about matching errors and the unethical use of face recognition technologies.

In summary

Face recognition technology vendors have a responsibility to implement best practices that identify and minimize any hidden biases, establish metrics for fairness, and test algorithms in real-world scenarios.

In recent years, the scientific community has been working together to improve training procedures, data and outcomes that reduce misidentifications not only based on gender, but also on ethnicity and other variables.

Know your algorithm! With more than 200 matching algorithms available, it is simply wrong to draw generalizations about algorithm performance overall and demographic effects in particular, especially when evaluating the technologies in a test environment. System owners should measure operational algorithm accuracy by conducting a proof-of-concept, or using a biometrics testing laboratory.

The risks hidden within automated face recognition processes, in particular irregular error rates for certain demographic groups, are well known, well documented and well argued. And governments worldwide continue to struggle with providing sensible regulations and standards that set definite rules for face recognition applications.

First and foremost, Cognitec strives to develop well-balanced algorithms that fairly match images for diverse, real-world populations. The company also contributes mindful expertise for establishing clear guidelines and performance expectations, which, in the end, foster the responsible and ethical use of face recognition technologies.

Cognitec

The trusted face recognition company since 2002