REGULATING FACE RECOGNITION TECHNOLOGY

WHY TESTING FACE RECOGNITION TECHNOLOGY IS CRITICAL—AND WHAT IS NEEDED

This brief explains what types of testing are necessary for meaningful evaluation of face recognition technology (FRT) for the most common law enforcement use: attempting to identify a witness, victim, or person suspected of committing a crime from an image—"investigative face identification." It is intended to aid legislators seeking to regulate police use of FRT.

KEY TAKEAWAYS

- FRT must be tested to know how well or poorly it works
- Current testing—including benchmark testing conducted by the National Institute of Standards and Technology (NIST)—is insufficient to evaluate the technology for law enforcement use
- FRT systems are complex human-machine systems—different types of testing throughout development and use are required to evaluate them fully
- Operational testing—assessing a system as it is actually used in the real world—is the best way to evaluate the accuracy and potential bias of a policing agency's particular FRT system

THE CONTEXT

Many states and localities are considering regulating police use of face recognition technology (FRT). If legislative bodies are going to permit policing agencies to continue to use—or start using—this technology, this use must be subject to carefully considered regulatory guardrails. Legislation should ensure that agencies only may use FRT if it makes the public safer and does not violate our fundamental rights and exacerbate harms like racial disparities in the criminal legal system.

Key to determining whether FRT makes the public safer is knowing how well or poorly the technology works. Put simply, it must be tested.

Unfortunately, testing of FRT as used by police today either is nonexistent or woefully inadequate. As a result, both lawmakers and the public lack the basic information needed to evaluate FRT's effectiveness and impact on public safety.

Below we describe the different testing required for FRT systems, what each type can (and can't) tell us, and how this testing should be incorporated into regulation.

Although testing is a necessary part of understanding FRT's impact on public safety, it is insufficient on its own. For more detail on the reporting and auditing requirements needed to evaluate the full impact of FRT use, see our Legislative Checklist.

EVALUATING FRT

International standards and best practices recommend three types of testing for FRT systems—each serving a distinct purpose: (1) technology testing to assess how well the FRT algorithm performs; (2) scenario testing to simulate a real-world use case; and (3) operational testing to assesses an FRT system as it is actually deployed in the real world.

As a useful analogue, imagine the testing required for another human-machine system: race cars.

When developing a race car, engineers first evaluate the mechanical components of the car (e.g., the engine and tires) → this is **technology testing**.

Once the mechanical parts have been tested, a trained driver test drives the car on a racetrack to predict how the car might perform in an actual race → this is **scenario testing**.

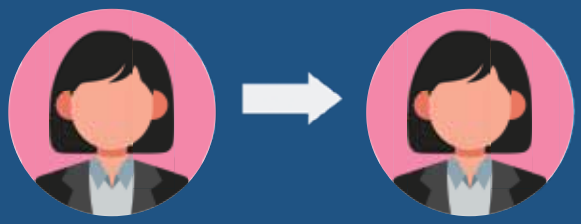
Finally, the race car is debuted in competition, where the car and actual driver's performance are measured in the field → this is **operational testing**.

FRT BASICS

WHAT IS FACE RECOGNITION TECHNOLOGY AND HOW DOES IT WORK?

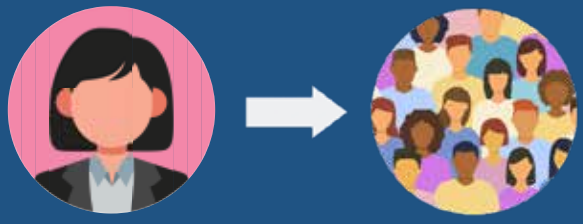
Face recognition technology uses computer algorithms to attempt to identify or verify the identity of a person or persons.

There are two main types of face recognition relevant to law enforcement use:



1. Verification—when an algorithm compares two images to try to answer the question, "is that you?" This is also called 1:1 matching because a "probe image"—the image inputted into a face recognition system—is only compared to one other stored image. (Think: iPhone's FaceID).

2. Identification—when an algorithm compares an unknown probe image to a database of images—the enrollment database—to attempt to answer the question, "who are you?" This is often called one-to-many matching (1:N) because it compares a probe image to all images in the enrollment database.



HOW DO POLICE USE FRT?

Police most commonly use FRT for face identification—to help identify an individual associated with a crime. A typical process might look like:

The actual process of using FRT requires human-machine interaction. A typical process might look like the following:

- An officer obtains a probe image of someone they want to identify. Common law enforcement probe images include cell phone images, video surveillance stills, such as from CCTV and ATM cameras, and social media images.
- The officer submits the probe image to the FRT software which compares the probe to the images in the enrollment database and returns a list of possible matches. Enrollment databases typically consist of mugshots, DMV images, and/or corrections department records (i.e., booking photos).
- An officer(s) then reviews the computer's possible matches to confirm an investigative lead.

1. TECHNOLOGY TESTING

WHAT IS IT?

Technology testing evaluates FRT algorithms, measuring criteria like error rates (both false positives and false negatives) and speed. It's like race car engineers testing the raw capabilities of a car's engine.

Currently, the National Institute of Standards and Technology (NIST) conducts field-leading technology tests of face recognition algorithms, assessing algorithmic accuracy and speed on its own private image databases under controlled, laboratory conditions.

WHAT IS NIST?

The National Institute of Standards and Technology (NIST), housed in the U.S. Department of Commerce, is the nation's leading physical sciences laboratory. Its mission is "to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology."

For the past two decades, NIST has led federal government efforts to develop standards for emergent biometrics and artificial intelligence technologies. As part of this work, NIST has created benchmark technical evaluations for face recognition algorithms. Its expertise and experience in this area is unrivaled.

WHAT IS IT USEFUL FOR?

Technology testing can establish a baseline of technical performance. Because NIST's tests compare algorithmic accuracy and efficiency *across* vendors, they spur technical improvement through competition and provide users with a yardstick to distinguish among algorithms. Applied to our race car analogy, it's akin to assessing how powerful a car's engine is.

WHAT ISN'T IT USEFUL FOR?

Technology testing can't tell you how an FRT system will perform in the real world.

There are two main reasons for this: (1) this testing doesn't evaluate the impact of the human reviewer → it doesn't test the race car driver; and (2) it doesn't evaluate algorithms on the images actually searched and the environmental conditions actually faced in the real-world → it doesn't test the race car and driver on the competition racetrack. Just as the ultimate test for a race car requires assessing the performance of the race car and driver in actual race conditions, FRT systems also must be assessed in their real-world use environments.

HOW SHOULD TECHNOLOGY TESTING BE INCORPORATED INTO LEGISLATION?

NIST's technology tests should serve a gatekeeping function—establishing floors that vendors must meet. Specifically, legislation should require that agencies only purchase FRT from a vendor that has demonstrated high accuracy—across relevant demographic groups—on NIST's technology tests.

Because technology testing can't tell us how a particular agency's FRT will perform in the real world, additional testing is required.

2. SCENARIO TESTING

WHAT IS IT?

Scenario testing evaluates entire face recognition systems—not just algorithms—in a laboratory setting designed to mimic a real-world use. It is a live, controlled experiment for a particular use case. Conducting a scenario test typically requires recruiting human volunteers to create test probe and enrollment image databases that closely resemble datasets in the intended use context. To use our race car comparison, scenario testing is like evaluating the car and test driver on a test racetrack.

WHAT IS IT USEFUL FOR?

Scenario testing can help predict full FRT system performance—rather than just the algorithm's raw technical capability—for an intended use case. Just like a test drive, scenario testing provides insight into potential sources of error in real-world use, identifying risks that developers and system owners should mitigate before the technology is actually deployed.

Legislative bodies or agencies could use scenario testing results to inform purchasing decisions—serving as a much more useful filter than NIST's technology tests. Skipping scenario testing of FRT systems would be akin to just testing the engine of a race car before debuting the full car in competition.

In the law enforcement context, scenario testing could help researchers and developers gain insight into potential vulnerabilities and sources of error for law enforcement uses and enable them to modify the technology accordingly. With access to scenario test results for different vendors, legislative bodies or agencies could use these results to inform acquisition decisions—serving as a much more effective filter than NIST's technology tests. Skipping scenario testing of FRT systems would be akin to just testing the engine of a race car before debuting the full car in competition.

HOW SHOULD THIS TESTING BE INCORPORATED INTO LEGISLATION?

Because scenario testing tends to be expensive and resource intensive, it is more feasible for a single statewide or federal agency to run these tests rather than individual policing agencies. Accuracy and bias information gained from a national or statewide scenario testing program would help narrow the field of vendors eligible for smaller agencies' use and provide vital information to industry for how to improve their technology for law enforcement use cases.

At the federal level, legislation could direct NIST, in collaboration with federal law enforcement, to develop a scenario testing program that models common law enforcement uses. Alternately, Congress could require that the Department of Justice commission a qualified biometrics testing lab to design a scenario testing program for law enforcement. (The Department of Homeland Security already employs this model at its Maryland Test Facility to assess biometric technology for travel identity verification.)

At the state level, a single state agency could be responsible for developing a scenario testing program in conjunction with an independent, qualified biometrics testing lab.

ROOTING OUT BIAS—SCENARIO TESTING IN ACTION

In 2020, during the peak of the COVID-19 pandemic, the Department of Homeland Security sponsored a scenario test to assess face masks' impact on the accuracy of FRT systems used for verifying traveler identity. This scenario test demonstrated that error rates for masked faces were higher, on average, for people with darker skin. This finding alerted vendors that they needed to update their FRT systems to address these racial disparities.

Without scenario testing, these racial disparities could have affected real people in deployment.

3. OPERATIONAL TESTING

WHAT IS IT?

Operational testing assesses an FRT system as it actually is used in the world—on the types and quality of images actually searched, the size of enrollment database used, and with a human reviewer evaluating the results.

WHAT IS IT USEFUL FOR?

Operational testing is the best way to know how well or poorly a particular FRT system actually performs.

The accuracy and bias of an agency's particular FRT system depend on several key contextual factors, including:

1. Image quality and demographic makeup of the images actually searched;
2. Size of the enrollment image database;
3. Product settings a user employs in the real-world (for example, the number of possible matches the system is set up to return); and
4. Decisions made by the human reviewers.

Proponents of police use of FRT often argue that operational testing is unnecessary because NIST's testing combined with a "human-in-the-loop"—the officer who reviews the computer's results—is adequate.

This argument is appealing but wrong:

- Algorithmic accuracy greatly depends on the type and quality of the images being searched and the demographic composition of the enrollment database—and NIST doesn't test algorithms on the probes images or enrollment databases that agencies actually are using. Most of the probe images used in NIST testing are much higher quality than the images commonly used by law enforcement, i.e., video surveillance images.
- Human reviewers can make things worse: cognitive biases are known to impact human interaction with machines. For example, humans may over-trust machine output—a phenomenon known as "automation bias." We also experience "other-race effect," a natural tendency to recognize faces of people of our own race more accurately than those belonging to different races. Race aside, research also shows that humans are innately bad at identifying unfamiliar faces—the exact task officers are asked to perform in reviewing FRT results. These cognitive biases mean we can't assume, without evaluating, that human reviewers improve the accuracy of the FRT process.

IMPACT OF IMAGE QUALITY ON ACCURACY

In its technology tests, NIST assesses probe images of varying quality. The takeaway is clear: lower image quality produces higher error rates.

For example: When analyzing a low-quality probe image—think computer webcam quality—the best algorithm makes 7x more errors than it does when analyzing high-quality images.

Why it matters: research and reporting have revealed that policing agencies have used a variety of low-quality images in FRT searches—from grainy surveillance camera photos to artist sketches. NIST testing shows that even the best algorithms will have much higher error rates on these types of images. Without regulation in place, police are free to use low-quality images regardless of these large error rates.

HOW SHOULD THIS TESTING BE INCORPORATED INTO LEGISLATION?

Any agency that uses FRT should be required to conduct operational testing or submit its system to testing by independent, third-party experts.

Ideally, if agencies are conducting their own testing, they should be required to follow a standardized, expert-developed testing protocol. A consensus testing protocol would enable standardized evaluation across state and local agencies. Unfortunately, nothing like this exists—yet.

Federal legislation could help fill this gap by directing and empowering an agency like NIST to develop an operational testing protocol that agencies could follow.

Because operational testing can analyze both the images actually searched and also evaluate the impact of the human reviewer, it is the only way to know real-word accuracy.

In the absence of a federally developed protocol, states must step up. Any state authorizing the use of FRT either should commission a task force of diverse, qualified experts to develop an operational testing protocol, or require that agencies engage independent, expert third-party testers to conduct this testing.