BIOMETRIC FACE RECOGNITION:
REFERENCES FOR POLICYMAKERS

AN INFORMATIONAL DOCUMENT CREATED BY THE FEDID COMMUNITY

Version 1.0 - December 2020

BIOMETRIC FACE RECOGNITION: REFERENCES FOR POLICYMAKERS

An informational document created by the FedID community

This is a community-developed reference document intended to provide value to policymakers[1] so they can gain an accurate, clear understanding of face recognition and its associated issues, thus allowing them to place inputs received into proper context. It neither advocates for or against the technology, for or against any application or use case, or for or against any legislative or policy outcomes. Furthermore, the document does not purport to represent the official policy or position of the federal government or any of its individual departments or agencies, nor of other government or private-sector entities

TABLE OF CONTENTS

This paper first briefly provides basic information about face recognition technology and common issues, so that all policymakers can accurately understand and assess the recommendations that they receive on the use of the technology. A second section provides more detailed information and insights that will be critical to policymakers who are crafting legislative or policy drafts, and the appendix provides more depth on a number of issues.

[1] For the purposes of this document, “policymakers” consist of executive and legislative personnel at any level of government.

INTRODUCTION

Face recognition (also referred to as facial recognition)[2] is one of the most powerful and complex technologies in modern times. Like many forms of intelligent systems, it is a tool that can be used for good or lead to harm, and it is incumbent upon each organization to leverage that tool responsibly.

This technology and its related concerns (such as privacy, error rates, and demographic differentials) have recently generated significant attention throughout the policy community, which is a positive development. By its very nature, face recognition will never exist without legitimate associated concerns, so its use must include the appropriate safeguards and strong privacy protections from the beginning. Unfortunately, policymaker efforts have often not focused on realistic situations or priority issues, since many aspects of this technology are difficult to grasp – even by those providing guidance and recommendations to the policymakers.

This paper, a volunteer-developed product of the FedID community[3], offers U.S. policymakers a clear understanding of face recognition technology. It neither advocates for nor against the technology, for nor against any application or use case, or for nor against any legislative or policy outcomes. Similarly, it does not delve into any individual application nor discuss existing policies and procedures that federal agencies are already using. Rather, it provides foundational insights on aspects and issues of face recognition relevant to policymakers, so that their subsequent investigations and discussions on those topics can be proper and productive.

[2] Throughout this paper, “face recognition” refers to biometric face recognition. Facial analytics are a different, though sometimes connected, technology. A more detailed explanation of the two is provided later in the paper.

[3] The federal identity (FedID) community includes public and private sector developers, evaluators, system owners, and operators that are engaged in the responsible and appropriate deployment and use of identity technologies by federal agencies.

BIOMETRICS AND FACE RECOGNITION 101

Face recognition is one of several different types of biometric modalities, all of which attempt to recognize the identity of an individual. Other examples include fingerprint, iris, and speaker recognition. It is important to understand the difference between (biometric) face recognition and other artificial intelligence (AI)-based algorithms that also use face images but perform different types of analysis (such as estimating age, gender, or ethnicity).

Not only do these technologies seek to perform different functions, but they also have different backgrounds, with biometrics having a long history of best practices and international standards guiding related activities. This section provides basic information on biometric technology in general and issues specific to the face recognition modality, and further differentiates (biometric) face recognition from other facial analytics algorithms.

BIOMETRIC BASICS

At a high level, the basic biometric process entails:

1. Gathering an observation (for face recognition, a photo of the individual's face)
2. Converting that observation into a biometric template[4] for use by a recognition algorithm
3. Comparing that template to one or more previously generated and stored templates, producing a similarity score[5] for each comparison
4. Comparing the similarity score to a user-selected threshold setting[6] and providing results of this comparison to system operators

This process is enabled by multiple components within a biometric system:

- Sensors, which read relevant information from an individual and convert it into a digital form. This often involves some type of user interface, liveness detection capability, and quality check.
- Template creation algorithms, which convert the individual's information into a biometric template for use by the recognition algorithm.
- Database, which provides one or more previously developed templates for comparisons.
- Recognition algorithm, which compares the new template to one or more from the database and creates a similarity score for each comparison.
- Decision rules, which determine the system's output. While they are fundamentally based on the comparison of the similarity score (determined by the recognition algorithm) and the threshold setting (set by the system administrator), they are in most cases only one of multiple factors that will be considered.
- In many, but not all, applications, human adjudicators, who determine recommended actions based on output review and additional non-biometric factors.

Template creation and recognition algorithms are the most fundamental and often-discussed components of a biometric system[7], with the two usually provided together within a vendor's packaged product[8]. Other components are often selected and managed by system operators, but all of them must be selected and tuned, individually and collectively, for each operational application. The performance of each component impacts the ability of the other components to do their jobs, with the latter components having to include measures to detect and accommodate non-ideal inputs.

THE EVOLUTION OF FACE RECOGNITION

The first semi-automated algorithm for face recognition was developed in the 1960s and required an administrator to locate features (such as eyes, ears, nose, and mouth) on the photographs before it calculated distances and ratios to a common reference point.

Reliable real-time automated face recognition began in the 1990s using the eigenfaces technique, which uses data compression to reveal low dimensional structures of facial patterns. (see image below) This approach, as well as Linear Discriminant Analysis and Elastic Bunch Graph Matching, dominated the market for a couple of decades.

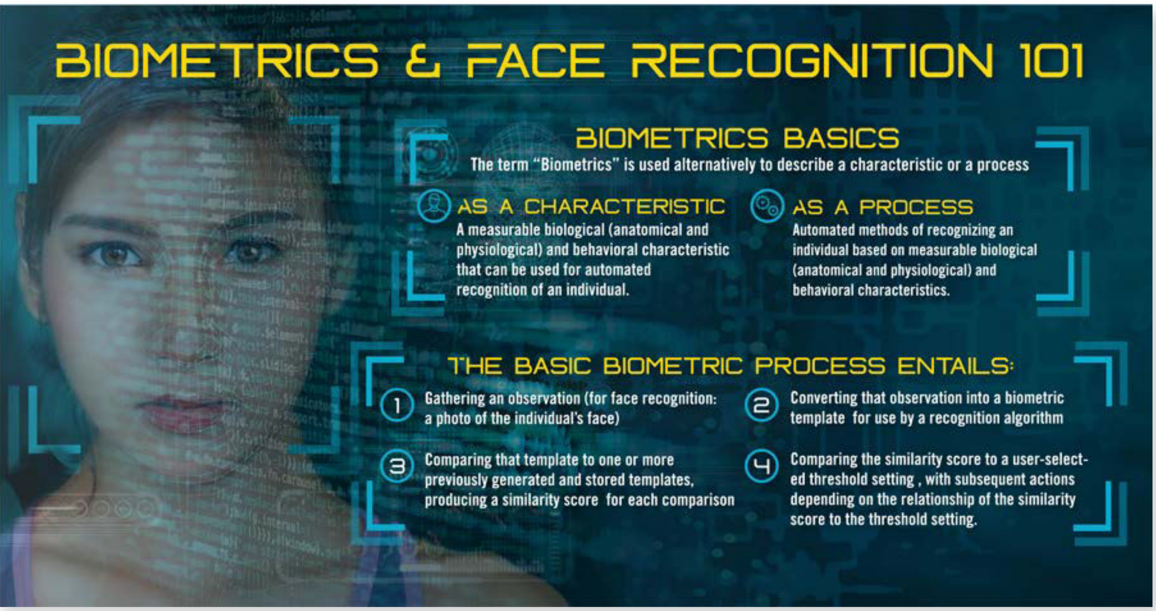
Modern face recognition algorithms are leveraging cutting-edge artificial intelligence to enhance or create new approaches to face recognition, which has lowered error rates considerably over the past few years.

Adapted from *Face Recognition*, a 2006 publication of the National Science and Technology Council.

[4] Template: A digital representation of an individual's distinct characteristics, representing information extracted from a biometric sample. Biometric templates are what are compared in a biometric recognition system. Source: *Biometrics Glossary,* a 2006 publication of the National Science and Technology Council.

[5] Similarity (or match) score: A value returned by a biometric algorithm that indicates the degree of similarity or correlation between a biometric sample and a reference template. Ibid.

[6] Threshold: A user setting for biometric systems leading to a predetermined workflow decision. The acceptance or rejection of biometric data is dependent on the similarity score falling above or below the threshold. The threshold is adjustable so that the biometric system can be more or less strict, depending on the requirements of the application



BIOMETRICS & FACE RECOGNITION 101

BIOMETRICS BASICS

The term "Biometrics" is used alternatively to describe a characteristic or a process

AS A CHARACTERISTIC

A measurable biological (anatomical and physiological) and behavioral characteristic that can be used for automated recognition of an individual.

AS A PROCESS

Automated methods of recognizing an individual based on measurable biological (anatomical and physiological) and behavioral characteristics.

THE BASIC BIOMETRIC PROCESS ENTAILS:

1 Gathering an observation (for face recognition: a photo of the individual's face)

2 Converting that observation into a biometric template for use by a recognition algorithm

3 Comparing that template to one or more previously generated and stored templates, producing a similarity score for each comparison

4 Comparing the similarity score to a user-selected threshold setting, with subsequent actions depending on the relationship of the similarity score to the threshold setting.

MEASURING BIOMETRIC PERFORMANCE

Measuring the performance of entire biometric systems, or the algorithmic subcomponents (which are two entirely different problems) in a nonbiased and statistically significant manner is quite complicated and costly. Issues that may at first seem inconsequential can have significant ramifications, leading to incorrect results. National and international standards for biometric performance testing and reporting[9] should be followed. The reliability of results from evaluations that do not follow these standards is highly suspect.

Modern biometric recognition algorithms have exceptionally low error rates in ideal conditions[10], but multiple factors can negatively influence those rates. The often-asked question, "How accurate is it?" cannot be validly answered with a single number. If only it were that easy! As biometric recognition algorithms are inherently probabilistic, it is both more proper and more insightful to think in terms of error rates. The proper error metric to use depends on the intended mode of operations: verification or identification (see box at right).

Verification

Consider a verification application. Ideally, the similarity scores for the correct user will always be higher than the threshold setting, and similarity scores for incorrect users (i.e., imposters) will always be lower than the threshold setting. If the algorithm produces a similarity score for the correct individual that is lower than the threshold setting, then the system has incorrectly rejected the individual. (John claims to be John, but the algorithm disagrees.) How often that occurs is the false reject rate (FRR). If the algorithm produces a similarity score for an imposter that is higher than the threshold setting, then the system has incorrectly accepted the wrong individual. (Alice claims to be Eve, and the algorithm agrees.) How often that occurs is the false accept rate (FAR). The two rates are connected by the threshold setting. As a system operator varies the threshold setting, both rates will change. The results are often mapped by using a chart like that shown in Figure 1.

Performance of different algorithms on a given database is often compared by mapping these curves for each algorithm on the same chart, such as Figure 2 (from a 2006 evaluation)[11].

VERIFICATION is a task where the biometric system attempts to confirm an individual's claimed identity by comparing a submitted sample to a previously enrolled template. (It is sometimes referred to as a "1:1 comparison," as the recognition algorithm is comparing the newly developed template to the previously developed template of the claimed individual.) One example is using biometrics to unlock your phone or computer.

IDENTIFICATION is a task where the biometric system attempts to determine the identity of an individual by comparing the new template to several existing templates, and rank-ordering the similarity scores. (It is sometimes referred to as a "1: many comparison," as the recognition algorithm is comparing the newly developed template to multiple previously developed templates from its database.) One example is comparing an arrestee to a criminal mug shot database

Note that recognition is a generic term and does not necessarily imply either verification or identification.

[7] This is definitely true for the face recognition biometric modality, and thus receives the majority of attention throughout this paper.

[8] Sensors may also be provided as part of a biometric commercial product, though this often is not the case for face recognition.

[9] The National Institute of Standards and Technology (NIST) manages the "Registry of U.S. Recommended Biometric Standards" per the NSTC's Policy for Enabling the Development, Adoption, and Use of Biometric Standards. Federal agencies can generally obtain free access to standards included in this registry. See https://www.nist.gov/itl/iad/image-group/support-registry-us-recommended-biometric-standards for additional information.

[10] For face recognition, think of a passport photo: high resolution, uniform lighting, constant and uncluttered background, and straight-on angle.

An alternative display approach that is often used is to select a FAR (say, of 0.01) and then state (in text or a table) what the resulting FRR would be for each algorithm at that consistently selected FAR. Stating a FAR or FRR without the corresponding rate does not provide insight into the algorithm's performance.

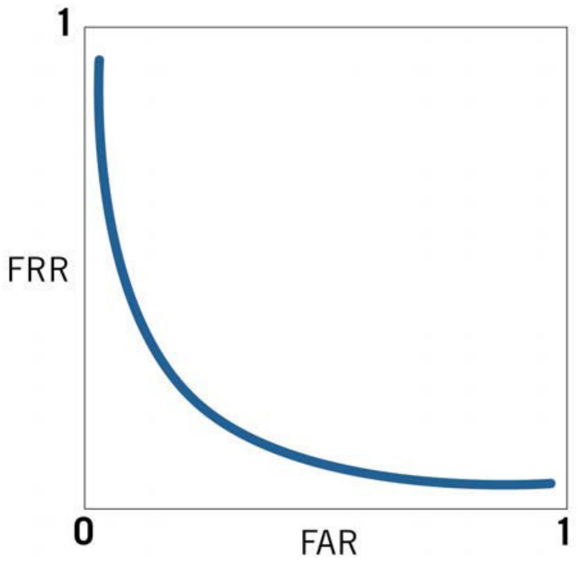


Figure 1.

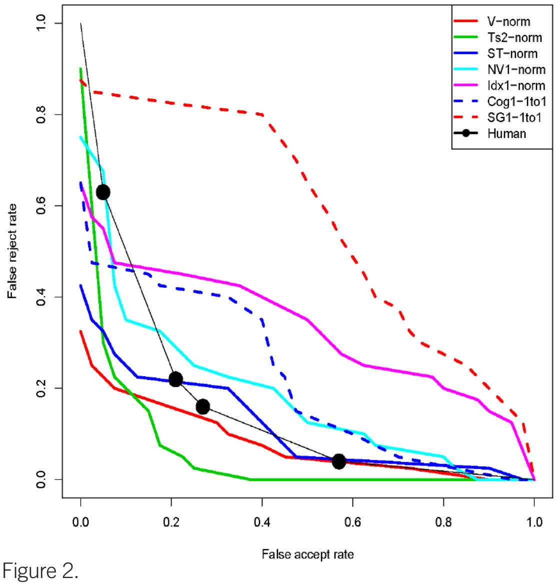


Figure 2.

Identification

Now consider an identification application, where the template of an individual is compared to numerous existing templates, and the similarity scores are rank-ordered. Ideally, the similarity score for the correct individual would be the highest (and for applications that also use a threshold, would be the only similarity score above the threshold). In reality, the correct match could have something like the fourth highest similarity score, and the top 10 scores could all be above the threshold. A graphical representation of measured identity performance is thus quite complicated: it would have two dimensions showing performance as the threshold setting varies (somewhat similar to those shown above for verification), but would have a third dimension to show the ranking perspective: Is it the top match? Or within the first and second ranked match? Or within the first, second, and third ranked match, and so on.

For simplicity's sake, many evaluators focus on the top match only, which will result in graphics that appear very similar to the verification chart shown above. While this is useful to determine directional trends for future research purposes, it is not entirely useful as an indicator for operational performance, as most identification applications require additional rank-ordered matches.

Statistical Significance and Other Evaluation Considerations

For both verification and identification statistics, the fidelity of these measurements depends on the number of individuals used and comparisons made, per standard statistics rules. Evaluations with higher numbers of individuals and comparisons will provide more precise results. Evaluations with only a few dozen individuals or comparisons will have high error variances, making their measurements (and any analysis based on them)

[11] Jonathon P. Phillips et al. *FRVT 2006 and ICE 2006 Large-Scale Results.* 2007. NIST, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=51131.

suspect. Most biometric modalities, including face recognition, have such low error rates that evaluations must have massive numbers of test subjects and comparisons to reach statistical significance.

The prior discussion is a high-level overview of measuring the performance of recognition algorithms only. However, the recognition algorithm is only one of many components in a biometric system, and the performance of each component impacts the capabilities and responsibilities of the components that work later in the system's process stream. System-level variances (such as changing acquisition sensors or even sensor settings) and even natural variances (such as gender, ancestry, or age of the individual) will also cause error rates to change. The final decision process must understand and address these impacts, both individually and collectively, to ensure that the system produces an optimal output.

Measured error rates will vary, sometimes significantly, across different vendors. Evaluations show a range of performance, with significant differences in accuracy between the best and worst performing vendors. Blanket statements about all algorithms based on an assessment of a subset will often be incorrect

FACE RECOGNITION-SPECIFIC INSIGHTS

Face recognition performance variances. Face recognition uses an image of the visible physical structure of an individual's face for recognition purposes. In addition to the typical biometric performance variances, face recognition will also vary based on factors such as lighting, pose angle, image quality, age of the individual, time difference between images being compared, and partial occlusion of the face (e.g., wearing a COVID-19 mask). In some face recognition applications, these factors can be limited (e.g. having fixed lighting and camera placement when taking drivers license photographs), while in others they cannot (e.g. in surveillance video, it is not possible to make sure each passer-by looks directly into the camera). Therefore less constrained applications will typically have worse face recognition accuracy.

Forensic anthropology has shown that there are naturally occurring facial variances across different demographic groups. These make it more difficult, but not impossible, to ensure similar face recognition algorithm error rates across demographic groups. This is an active area of current research, with NIST's recent Face Recognition Vendor Test (FRVT) demographic report[12] providing valuable insights for future research.

The variances measured in the assessment of recognition algorithms (only) do not automatically imply that equivalent variances will always occur in the output of operational systems, as the algorithm is one component of a complex human-machine operational system. Other system components should be designed to handle these, and additional, variances.

Biometrics vs. analytics. System owners use face recognition to verify or attempt to determine the identity of an individual. Face recognition does not explicitly attempt to determine the gender, age, or ethnicity, or to recognize the facial expression or medical conditions, of an individual. These are functions of a different category of AI-based capabilities, best classified as facial analytics algorithms, which do not attempt to determine the identity of individuals.

Systems can use both face recognition and facial analytics algorithms in combination to enhance overall system performance. However, conflating error rates and associated issues of face recognition with facial analytics algorithms will usually lead to inaccurate analyses and misleading face recognition policy conclusions.

TIPS FOR THE POLICYMAKERS

Developing assumptions about operational biometric system performance, or impacts across different demographic groups, from recognition algorithm performance metrics only is improper, often leading to inaccurate conclusions.

Face recognition is an example of an "emergent" system, where the system s behavior is a consequence of the interactions and relationships amongst its components, rather than the independent behavior of individual elements. Evaluating an operational system s performance thus requires an end to end analysis.

[12] Patrick Grother et al. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects.* 2019. NIST, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf

FACE RECOGNITION

Is this the same person?

FACIAL ANALYTICS

Estimate subject's age, gender or ethnicity.

Determine their facial expression.

Look for medical conditions.

Application variances. Face recognition has many different applications, where issues such as error rates, demographic performance, privacy and civil liberties, ethical appropriateness, and management and oversight requirements will vary substantially. Policy deliberations should therefore be specific to an individual application (or a use case category that consists of multiple similar applications). General investigations or improper mixing of multiple use cases will lead to incorrect assessments and decisions.

Biometric vs. human recognition abilities. Since 2006 (See Figure 2), researchers have been able to show that the best face recognition algorithms were more accurate than untrained humans at deciding whether two images "matched" or not. Although humans are very good at recognizing people with whom they are familiar (e.g., family, friends, celebrities), the average person is not very good at this task when they are unfamiliar with the subjects. The gap in performance between the layman and face recognition algorithms has only widened in the last 15 years as algorithms have become better and better.

This is not to say that some humans are not very good at this task. Recent research[13] has demonstrated that trained professionals who perform 1:1 facial comparisons or adjudicate candidate lists are as good as the latest face recognition algorithms, and this "expert human" level of performance is on a par with the accuracy shown by expert fingerprint examiners when they make decisions about latent fingerprint comparisons. Face recognition algorithms, however, can make these comparisons significantly faster.

This same research also supports the operational fusion of face recognition systems with human adjudicators to achieve the highest accuracy when performing face recognition searches. In other words, to achieve the highest level of accuracy, the current science indicates that algorithms should be combined with expert humans.[14]

[13] P. Jonathon Phillips et al. *Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms.* 2018. Proceedings of the National Academy of Sciences of the United States of America, https://www.pnas.org/content/pnas/115/24/6171.full.pdf.

See also Kimberly Underwood. *The Accuracy of Machines in Facial Recognition 2020.* AFCEA, https://www.afcea.org/content/node/22559/. Accessed September 9, 2020.

[14] This is how most agencies use face recognition today when performing investigative searches – the algorithm returns a candidate list, which is then adjudicated by a human to determine if any candidate is worth flagging as an investigative lead.

FACE RECOGNITION: AREAS FOR POTENTIAL POLICYMAKER CONSIDERATION

INTRODUCTION

There are many areas within face recognition for possible policymaker attention. While attention on any individual area could be beneficial, it must be done while recognizing the influences from, and impacts to, other areas. Policymakers must therefore maintain a comprehensive viewpoint even while addressing any individual area. For example, restricting researchers' access to a specific data set could be a wise decision for many different reasons, but doing so could also inhibit those researchers' ability to overcome an algorithm's differential performance issues across different demographic groups. A more balanced, mature approach may therefore be needed.

Policymakers must be concerned not only with the individual areas of activity within the face recognition community (such as research and development, testing, or system planning) but also with areas of common concern (such as data, privacy, and oversight). The table below provides a mental framework to understand the interplay between these areas. For example, while there will be some common data issues between the research and operational sustainment activity stages, there will also be data considerations that are specific to each stage.

	RESEARCH AND DEVELOPMENT	TESTING	INITIAL IMPLIMENTATION CONSIDERATIONS	SYSTEM DESIGN, POLICIES, AND PROCEDURES	OPERATIONAL SUSTAINMENT	DISPOSAL
DATA						
PRIVACY AND CIVIL LIBERTIES						
SAFEGUARDS AND OVERSIGHT						
REPORTING						

This section discusses each of the areas of activity (the columns in the table above) and areas of common concern (the rows), providing face recognition-specific insights for each. The appendix focuses on areas where additional insights are needed (such as data for research or privacy considerations when terminating an operational system).

Also recall from the Biometrics and Face Recognition 101 section that most face recognition policy deliberations must be application specific. Issues such as error rates, demographic performance, privacy and civil liberties, ethical appropriateness, and management and oversight requirements will vary substantially across different applications. There is also a wide variety of potential applications for face recognition, so attempting to list and describe all of them would make this paper unwieldy. Therefore, this paper has created three "use case categories"[15] where similar applications can be grouped and discussed at a high level. Doing so enables us to provide policymakers with critical insights and a starting point for their in-depth analyses.

Use Case Category 1: Verifying Claimed Identities

In this use case category, face recognition is being used to help verify a claimed identity (a verification application). The most common example is when individuals use the technology to unlock their phones. Other applications include verifying identities during international travel or supporting access control to secure facilities.

Use Case Category 2: Identity Determinations

In this use case category, face recognition is used to find potential matches in large databases (an identification application, usually without thresholds). The most common examples are by state departments of motor vehicles or national passport offices as they work to find fraudulent duplicates of identity credentials. Another use is when law enforcement works within specific investigations to determine the identities of individuals suspected of committing crimes. The face recognition system returns results (one or more rank-ordered matches, depending on the individual application) as a lead for additional human review and further investigation.

Use Case Category 3: Watchlist

In this use case category, authorities use face recognition to search for known threats at a specific location (an identification application, usually with thresholds). An example is when authorities search for known and suspected terrorists among a group of individuals congregating around a secure facility. The application's watchlist database is small and targeted to likely threats, based on specific legal and policy criteria. When a comparison produces a similarity score higher than the threshold, an alert of the potential match is provided to authorities for further investigation.

AREAS OF ACTIVITY

Policymakers need to focus on a range of activities throughout the face recognition continuum. Individual stages within this continuum have their own areas of concern that must be addressed, with an eye on how they will impact future stages. The following areas of activity are discussed:

- Research and Development
- Testing
- Initial Implementation Considerations
- System Design, Policies, and Procedures
- Sustainment
- Disposal

WHAT ABOUT THE USE CASE OF CONTINUAL MASS SURVEILLANCE?

The use case most often raised as a concern for face recognition is that of widespread, mass surveillance. The premise is that of an organization or entity in control of an interconnected network of cameras and face recognition systems that is capable of accurately identifying every individual it encounters and tracking their whereabouts. This use case is sometimes augmented with additional technologies that could understand what everyone is doing and compile that information as well. Fortunately, applications within this use case category fall within the realm of science fiction rather than reality. Current hurdles include:

- Error free face recognition. For such an application to function, error rates for face recognition systems would have to be virtually non existent, or else system operators would be overwhelmed with review and adjudication demands. Face recognition algorithms are highly capable, but nowhere near this level of accuracy. In fact, there is no current evidence that human faces themselves are sufficiently unique for this to be even theoretically possible.
- Significant collection of cameras. These applications would require a staggering number of high resolution cameras that would be capable of producing usable face images at any location in any condition, in real time.
- Infrastructure operations and maintenance. These applications would require massive network, database, and computational resources, which would have unmanageable fiscal and manpower requirements.
  - As an example, consider current estimates that Chicago has 32,000 closed circuit television cameras, which translates to ~768,000 hours of video per day. As high as these numbers are, the cameras do not cover the entire city, and not at a sufficient resolution to enable error-free face recognition.
- What database? There is no single, all encompassing face database in the United States. Only about 44% of the U.S. population (143 million) have passports. Individual states have their own driver license databases that are not connected to the passport system, nor to one another. Likewise, individual state and municipal criminal mug shot systems are not linked.
- Societal approval. The physical and fiscal requirements for these applications mean they could not be done in secret, and citizens of free and open democratic societies are loathe to accept this type of all encompassing monitoring.

Given these limitations, this paper does not consider applications within the continual mass surveillance use case to be currently feasible.

[15] These use case categories are not formal definitions within the face recognition community but were crafted to aid discussion within this paper.

RESEARCH AND DEVELOPMENT

Continuing research on face recognition algorithms, machine learning, camera technology, and computational resources is crucial for developing and deploying fast and accurate face recognition systems. Historical face recognition research has focused on both the general performance of individual algorithms and the larger system. In the future, enhanced research is expected to focus on more precise aspects of the performance of the algorithms, the impact of different inputs, and training approaches to building the algorithms.

The majority of face recognition research is sponsored and performed by the private sector, though it often leverages insights and new capabilities enabled through earlier (and more generic) government-sponsored research. The federal government sponsors targeted face recognition research to meet specialized needs or to advance the state of the art in directions that would not otherwise be possible or a priority for the private sector.

Although there are no legal requirements specific to face recognition research, generic requirements and best practices do exist and should be followed. For example:

- Face recognition research requires a significant amount of data (face images and necessary metadata about those images to enable the research, such as time and date of the images, who is depicted, etc.). Researchers must ensure that this data is legally obtained, their use for it is allowable (usually by informed consent of the individual or by the owner of a set of data), and access to the data is properly secured.
- Researchers are trained to follow a code of ethics specific to their field of study. Face recognition research often takes place at the intersection of multiple disciplines, however, requiring researchers to understand and follow multiple sets of best practices. (For example, AI researchers may not always have experience dealing with personally identifiable information.)
- Because of the sensitivity of face data, many research entities require their staff to subject their research and security plans to analysis and approval by their Institutional Review Board (IRB) before beginning the research.
- Most research results are protected using standard intellectual property procedures and reported in technical journals, so that the typical evolutionary cycle of research, where several researchers build on each other's prior discoveries to turn ideas into real-world technologies, continues.

TIPS FOR THE POLICYMAKERS

Policymakers' primary focus on face recognition research should be on the data – was it collected, used, and protected properly?

Policymakers can also influence researchers to prioritize striving to achieve equitable outcomes of algorithm metrics across various demographic groups.

TESTING

Face recognition algorithms and other system components must be extensively tested to identify areas of additional research, to inform decisions while planning operational systems, and to monitor operational performance. As algorithms and systems continue to improve, the complexity and specificity of testing will also have to increase. There are three different types of face recognition testing, each serving a different purpose. It is important for policymakers to understand the differences among the three, and how to consider their outputs.

- Technology Evaluations assess the abilities of face recognition algorithms only. They typically involve massive numbers of subjects in standard data sets so that performance variation across different algorithms can be measured and compared. Results from these evaluations are used to identify areas that require additional research or as a first step in selecting an algorithm for operational use. Highlighting any result from a technology evaluation and claiming that to be the expected outcome within an operational system will almost always be incorrect. The gold standard for face recognition technology evaluations is NIST's long-standing FRVT series [16].

THERE ARE HUNDREDS OF FACE RECOGNITION ALGORITHMS AVAILABLE TODAY, AND THEIR PERFORMANCE VARIES WIDELY. MOST STATE-OF-THE-ART ALGORITHMS ARE PROPRIETARY AND CAN BE TUNED FOR OPTIMAL PERFORMANCE IN A SPECIFIC APPLICATION. NO SINGLE ALGORITHM CAN EXCEL OR BE "BEST" FOR EVERY POTENTIAL APPLICATION.

[16] *Face Recognition Vendor Test (FRVT).* National Institute of Standards and Technology, https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt. Accessed: October 29, 2020.

- Scenario Evaluations enable initial assessments of how a system with a face recognition algorithm as a component will perform in a specific application. A mock-up of the anticipated operational environment is created, and humans are live subjects throughout the test. Scenario evaluations involving two different systems would have the same environment and subjects, but they would receive their own input data from the live subjects. To obtain maximum insight, the databases used in scenario evaluations need to be from "real" operational captures, thus necessitating policy-level authority for operational entities to share this data with approved testing entities. Results from scenario evaluations offer a good understanding of how the system will operate in the real world, thus providing potential operators input on selecting systems and establishing operational procedures. Scenario evaluation results are also specific to the application and the system tested. Highlighting results as being what would be expected for different systems in the same application, or the same system in other applications (or in general), will almost always be incorrect.

- Operational Evaluations are evaluations of a specific system in a specific use case at an actual operational location. They do not usually measure accuracy (though it can sometimes be feasible), but rather analyze other factors such as workflow impact and customer experience. Results from operational evaluations are typically used to enhance procedures within the operational system.

Data and subject selection are critically important and difficult for all evaluations. Any minor improper selection, or unobserved variations across data sets, can dramatically impact the tests, leading to inaccurate results. The numbers of test subjects and trials will also impact the accuracy of the results.

The organizations evaluating face recognition systems must meet standard requirements to ensure confidence in objective, impartial evaluation results.

More detailed insights on biometric performance testing and reporting can be found in the ISO/IEC 19795-1 standard[17]. Testing methodologies, laboratories, and evaluation results that do not follow these guidelines are highly suspect and generally should not be trusted.

INITIAL IMPLEMENTATION CONSIDERATIONS

Agencies beginning to plan an operational activity that includes face recognition need to answer many questions as they thoroughly analyze, decide, and document how the envisioned system will be used. Before jumping into that step, however, some fundamental concerns and issues must first be considered.

By its very nature, face recognition leverages personally identifiable information (PII), which creates ethical and privacy (as well as security) concerns. The first, and most important, question must therefore be, is the application ethically appropriate and legally permitted, and will privacy and security of this information be maintained? Additional considerations include:

- Has research and evaluation shown that leveraging face recognition in this application will provide the desired outcome?
- Do the benefits of leveraging face recognition in this application outweigh the concerns its usage creates?
- Does the conceptual model meet operational requirements while also maximizing protection of privacy and civil liberties?
- What notice, if any, will need to be given?
- Are there adequate resources and plans for training system operators, reviewing their work, and providing necessary oversight of the operational system?

TIPS FOR THE POLICYMAKERS

Information about face recognition evaluations must be sufficiently detailed to warrant consideration of their results. Policymakers first need to assess whether the evaluations were properly designed and are statistically significant, and whether the subjects and how they were presented enabled necessary repeatability without introducing prejudice or accidental influence. They must also understand specifics of the algorithm or system being evaluated under which contexts. Only then will policymakers be able to assess the results and know under which contexts those results will be meaningful.

[17] *Information technology — Biometric performance testing and reporting — Part 1: Principles and framework. International Organization for Standardization*, https://www.iso.org/standard/41447.html. Accessed October 29, 2020. (Note that this standard is currently being updated, with publishing anticipated in the next few months.)

Additional Study Guide

Additional Considerations for Use Case Category 1: Verifying Claimed Identities.

- This use case is more dependent on user collaboration than any other.
  - Is there an adequate plan to explain the purpose, risks, and benefits of the system to these users?
  - Are adequate training and help desk capabilities planned to support user needs throughout the lifecycle?
  - What are the initial and ongoing predictions for user adoption of the system?
- Will use of the application be mandatory, or will individual users be able to opt in or opt out?
- What are the safeguards or alternative approaches for users who cannot use the face recognition system?
- What is the impact of false accepts (i.e., the system accepts the claim of an imposter that they are someone else), and false rejects (i.e., the system denies the true claim of an individual that they are who they say they are)? How will they be handled?

Additional Considerations for Use Case Category 2: Identity Determinations.

- What is the role of human review(s) of outputs for this application? What is the review and oversight process for their work to ensure they use the outputs properly?
- What training and oversight will be required to ensure that operators understand what the outputs from the face recognition system mean?
- What will be required to document the circumstances of each analysis and result, and how those results were leveraged?
- What is the impact of incorrect or missed identifications? How will they be handled and explained?

Additional Considerations for Use Case Category 3: Watchlist.

The questions for use case Category 2 apply for this use case as well, though the deliberations on how to answer them, and thus the answers themselves, will be different for this use case. For example, this use case usually has a much more compressed time requirement, the implications for false alarms or missed identification can be much more extreme, and documentation and result retention needs will be vastly different. Additional considerations include:

- What are the criteria and processes for creating and maintaining the entities in the watchlist?
- What are the time performance objectives and how will they be achieved (e.g., how will "hits" be reviewed and handled in real time)?
- What information needs to be provided back to the entity that originally nominated the individual to be included on the watchlist?

TIPS FOR THE POLICYMAKERS

This area is ripe for policymaker influence, since guidance or use constrictions are lacking. As with most face recognition issues, specificity will be key:

- Which applications or use case categories should be prohibited or usually restricted, and how?
- Which applications or use case categories are generally acceptable, provided certain conditions are met?
- Which applications or use case categories fall into gray zones that require further analysis and oversight, and by whom?

SYSTEM DESIGN, POLICIES, AND PROCEDURES

The answers to the questions posed in the Initial Implementation Considerations stage serve as a foundation for designing the system and developing policies and procedures governing its operation. During this stage, planners use their preferred system lifecycle model, consistent with their current practice. For example, most federal agencies that are designing a face recognition application use a well-structured lifecycle model with the following components:

- Requirements
- Concept of operations
- Privacy and policy approvals
- System design
- Operational testing
- Sustainment
- Disposal

The first five of these elements are grouped in this section, with the latter two addressed separately in the sections immediately following. Policymakers should be mindful of the following issues while working through the operator's lifecycle planning process:

- Have potential operators analyzed all system components (individually and as an integrated system) to know which should be selected for this operational application?
- Have system operators optimized settings for this operational application?
- How will the output from the face recognition component of the system be used, along with what other information, to achieve desired outcomes?
- Is the plan to manage operational data, from both security and privacy perspectives, appropriate and sufficient?
- Which portions of the system will be fully automated, and which will require operator assistance or review?
- Is the overall system designed to enable optimal and suboptimal inputs to the face recognition component, and to accommodate anticipated errors from it?
- Will the system provide sufficient information to enable oversight review (e.g., auditing)?
- How will the system ensure appropriate outcomes across different demographic groups?

TIPS FOR THE POLICYMAKERS

Lack of rigorous planning for applications of face recognition systems and failure to follow established policies, rather than the technology itself, have been the root cause of many of the newsworthy "failures" of face recognition.

Policymaker assistance in ensuring proper planning and adherence to operational policies could be beneficial.

Additional Study Guide

Additional Considerations for Use Case Category 2: Identity Determinations.

- Has the appropriate data retention policy been developed to balance two equally important but opposing considerations?
  - To delete non-flagged data as rapidly as possible
  - To maintain records necessary to complete the investigation and support future prosecution
- Have adequate user supports been incorporated into the design to support desired user behavior and decision making? Example: does the output remind the operator that these are investigative leads and not a complete identity determination?
- Have agencies implemented procedures to confirm the effectiveness of operational systems (including adjudication) through proficiency testing (and auditing procedures) of human examiners?

Additional Considerations for Use Case Category 3: Watchlist

The questions for use case Category 2 apply for this use case as well, though the deliberations on how to answer them, and thus the answers themselves, will vary. Additional considerations include:

- Should the level of review and action determination vary based on the level of threat anticipated from the identified individual?

OPERATIONAL SUSTAINMENT

Sustainment generally includes activities related to ongoing operations and maintenance of the system, with typical issues such as maintaining system performance, patching and other system improvements, ensuring the system is operating effectively, and auditing. Additional concerns include:

- Is the plan to ensure security and data access management appropriate for the PII collected?
- Were unanticipated issues or changes in operational requirements adequately addressed?
- Can the system accommodate changes in procedure or interface to enable smoother operations?
- Are policies and system settings routinely analyzed to enhance system performance?
- Are the safeguards and oversight functions operating meaningfully?
- Is the delivered benefit sufficient to overcome the costs (fiscal, personnel resources, privacy impact)?

DISPOSAL

Face recognition systems cannot simply be shut down and forgotten. The data they use and create is PII and sensitive. While some of that data can be permanently deleted, other data may need continued storage and maintenance for future investigative or oversight purposes, depending on the application. Ideally, insights and lessons learned from the experience should also be recorded and shared for the benefit of future applications.

Disposal considerations tend to be specific to an individual area of activity. See the appendix for additional information.

TIPS FOR THE POLICYMAKERS

Policymakers focus within operational sustainment will be on adhering to policies and promoting continuous improvements, provided needed attention was paid throughout earlier areas of activity.

If prior areas were not properly managed, then all of those questions will resurface here – but it will be much more difficult to properly address them!

AREAS OF COMMON CONCERN

In addition to investigating face recognition areas of activity, several additional areas of concern must also be considered:

- Data
- Privacy and Civil Liberties
- Safeguards and Oversight
- Reporting

DATA

Data is the lifeblood of face recognition. It enables algorithm improvements, is the foundation for statistically significant evaluations, and provides the reference for template comparisons in operational settings. By the nature of face recognition, each data element is itself PII, or is at least connected to PII to enable system functionality. That means the data is not only a significant security risk but also raises privacy and civil liberties concerns.

Systems incorporating face recognition use many different types of data, including image data, match data (the results of matching two face images), and personal data that the system links to both individual files and to matches. Managing the data that is collected and generated by face recognition systems is critical to all aspects of security and privacy.

There are no U.S. laws or policies governing face recognition data specifically, though they do exist for PII. Nor are there overarching laws or policies that govern the use of sensitive data, though there are for data within some specific domains.[18] The Federal Data Strategy[19] provides 10 principles for federal use of data, including recommendations for ethical governance and conscious design. These are good high-level points of reference when investigating face recognition policy.

Biometric data can hold significant value, and significantly more value when it is connected to additional information. As a result, operators often have little incentive to delete face recognition data, unless specifically required to do so. Developing best practices or requirements for handling data in face recognition systems should be considered,[17] but these requirements must recognize that issues and considerations will vary across the system lifecycle (for example, research considerations are not the same as operational considerations).

Finally, recall from the Biometrics and Face Recognition 101 section that national and international standards (including for face recognition capture, quality assessment, and interchange) should be followed.

PRIVACY AND CIVIL LIBERTIES

Face recognition will always have privacy and civil liberties concerns that need to be addressed, within each of the areas previously described. While there are legitimate concerns with data sets used for research and testing purposes, most of these concerns are found in the operational stages. The proper time to begin developing privacy and civil liberties protections into operational face recognition programs is in the initial implementation consideration stage.

It is helpful to address the most significant privacy and civil liberties considerations for face recognition within the context of the Fair Information Practice Principles (FIPPs). The FIPPs are a set of internationally recognized principles that inform privacy policies within both the government and the private sector. These principles have been incorporated into data privacy laws, policies, and governance documents. For federal agencies, these principles have

TIPS FOR THE POLICYMAKERS

Those seeking guidance on how to manage face recognition data must cross-analyze numerous generic data and applications-specific laws and regulations, determining for themselves which are relevant and how to apply them. This can produce inconsistencies and non-ideal interpretations.

Developing clearer guidance and best practices, particularly for those without legal and regulatory experience, would be beneficial.

THE PROPER TIME TO BEGIN DEVELOPING PRIVACY AND CIVIL LIBERTIES PROTECTIONS INTO OPERATIONAL FACE RECOGNITION PROGRAMS IS AT THE VERY BEGINNING – IN THE INITIAL IMPLEMENTATION CONSIDERATION STAGE.

[18] Such as the Driver's Privacy Protection Act and the Health Insurance Portability and Accountability Act

[19] *Federal Data Strategy – Leveraging Data as a Strategic Asset.* Executive Office of the President, https://strategy.data.gov/. Accessed October 29, 2020.

[20] There are several laws in the European Union with regard to the need to completely delete biometrics data on the request of the subject, and laws that restrict government biometrics system owners from sharing biometric data outside of its original application. These could be informative while developing future U.S. legislation or policy.

been largely incorporated into the Privacy Act of 1974. For any legislators or policymakers considering face recognition mandates or guidance, the FIPPs provide a framework for privacy and civil liberties protections. The paragraphs below summarize the FIPPs in the context of face recognition.

Purpose Specification: Agencies collecting and sharing PII for face recognition testing, research, or operational use should specifically articulate the legal authority that permits such use of the face images and any associated data. They should ensure that a valid, lawful purpose exists for the collection and should identify the purpose for which the data was collected initially, and should try to limit subsequent use of the data to compatible uses.

Purpose specification should address the authority for collecting both the photos that constitute the photo database and any photos that are searched against the database. A specific challenge for face recognition is that photos are often collected for one purpose (e.g., to obtain a driver's license) and then are subject to face recognition for another purpose (e.g., law enforcement investigations). Agencies should establish clear policies to determine the most appropriate uses of face recognition and should enter agreements with providers of photos and/or users of their face recognition systems to ensure that all uses are permissible and data is securely protected.

Data Quality/Integrity: Agencies collecting and sharing PII for face recognition testing, research, or operational use should ensure that the face images and associated data are accurate, complete, and up-to-date. They should establish policies that safeguard the PII, update the PII whenever relevant new information is collected, and create a process for deleting data that is inaccurate or no longer needed.[21] Specific to face recognition systems, agencies should follow the guidelines and best practices for collecting/searching only those face images that meet data quality requirements and that will permit the face algorithm to return viable candidates to the users. Data quality is an especially important component in ensuring that the face recognition system does not misidentify individuals. Misidentification has been a significant concern regarding the implementation of face recognition systems, and agencies should consider additional ameliorative actions, such as limiting law enforcement action based solely on face recognition results and requiring specialized training of law enforcement.

Collection Limitation/Data Minimization: Agencies collecting and sharing PII for face recognition testing, research, or operational use should collect only those face images that are directly relevant and necessary to accomplish the specified purpose. The face images should be obtained by lawful and fair means and retained only as long as necessary to fulfill the specified purpose. Agencies may choose to accept only face images for retention or searching that meet a minimum legal threshold, such as reasonable suspicion or probable cause. They should be especially cautious to avoid the use of any face images that may have been collected in a manner that violates or chills an individual's Constitutional rights, such as the exercise of those rights guaranteed by the First Amendment.

Use Limitation/Security/Accountability: Agencies collecting and sharing PII for face recognition testing, research, or operational use should not disclose or make such data available except with the consent of the individual or by authority of the law. They should also follow stringent information technology and operational security to ensure only authorized access to the data.[22] Best practices may include justifications for the face recognition searches, logs of such searches, and frequent audits of the system. Agencies should institute reasonable security safeguards to protect the face images and associated data from unauthorized access, destruction, misuse, modification, or disclosure. They should also ensure that all employees receive training regarding privacy, security, PII breaches, and other relevant topics.

TIPS FOR THE POLICYMAKERS

The FIPPs provide a framework for helping to ensure privacy and civil liberties protections, as do PIAs and SORNs. The detailed analyses and decisions required to develop these documents help to ensure that closer attention is given to privacy considerations during system design but do not on their own guarantee optimal protection.

This issue is also far from limited to face recognition technology, as the nation's overall privacy construct is decades old. Technology innovation over the past few decades has far surpassed that of privacy protection. Policymaker attention to fixing that fundamental problem will create a stronger foundation for more advanced face recognition protections.

[21] This is an important aspect within the operational lifecycle planning activity. Many federal systems, for example, have continuous image quality and integrity processes built in.

[22] The major federal systems using face recognition are designed with access management and business rules that enable capabilities to limit access, use of data, auditing, etc.

Transparency: To the extent possible, agencies should be open about developments, practices, and policies for the collection, use, dissemination, and maintenance of PII for face recognition purposes. Although there is no federal legislation specific to face recognition, federal agencies must comply with both the Privacy Act of 1974 and the E-Government Act of 2002. Pursuant to these statutes, the agencies must publish both System of Records Notices (SORNs) and Privacy Impact Assessments (PIAs). Both SORNs and PIAs can provide significant transparency to the public regarding the use of face recognition and should be published in a timely and comprehensive manner. In addition, agencies may choose to publish policy guidance and engage with oversight bodies and advocacy groups to ensure an ongoing dialogue and accountability regarding their face recognition systems.

Individual Participation: To the extent practicable, agencies should involve the individual in the process of using PII and seek the individual's consent for the collection, use, dissemination, and maintenance of the face images. Although this may not be entirely possible in law or immigration enforcement applications, the agency can still collect the information directly from the individual when possible and provide access and redress rights to the individual. To ensure notice to the individual, the agency may need to resort to more public notices such as SORNs and PIAs.

SAFEGUARDS AND OVERSIGHT

Safeguards and oversight are critical for the operation of any system that deals with personal data and decisions that impact people's lives. In the case of face recognition, safeguards are needed to protect data and to ensure that data is used or shared only for the specific reasons that it was collected or generated, and (in most use cases) that the subjects of the data know how the data is being used and with whom it is being shared. In addition, it is important that the organizations that are providing oversight understand not only privacy and civil liberties but also face recognition and the applications in which face recognition is used.

Safeguards and oversight are generally very specific to an individual area of consideration. See the appendix for additional information.

REPORTING

Documenting procedures, results, and insights gained is a fundamental practice to enable others to understand and leverage one's work. Reports must be sufficiently complete for readers to understand the basis of the author's activities and to assess the accuracy and relevance of their work. Requirements for what is "complete" in documentation will vary based on the area of consideration.

TIPS FOR THE POLICYMAKERS

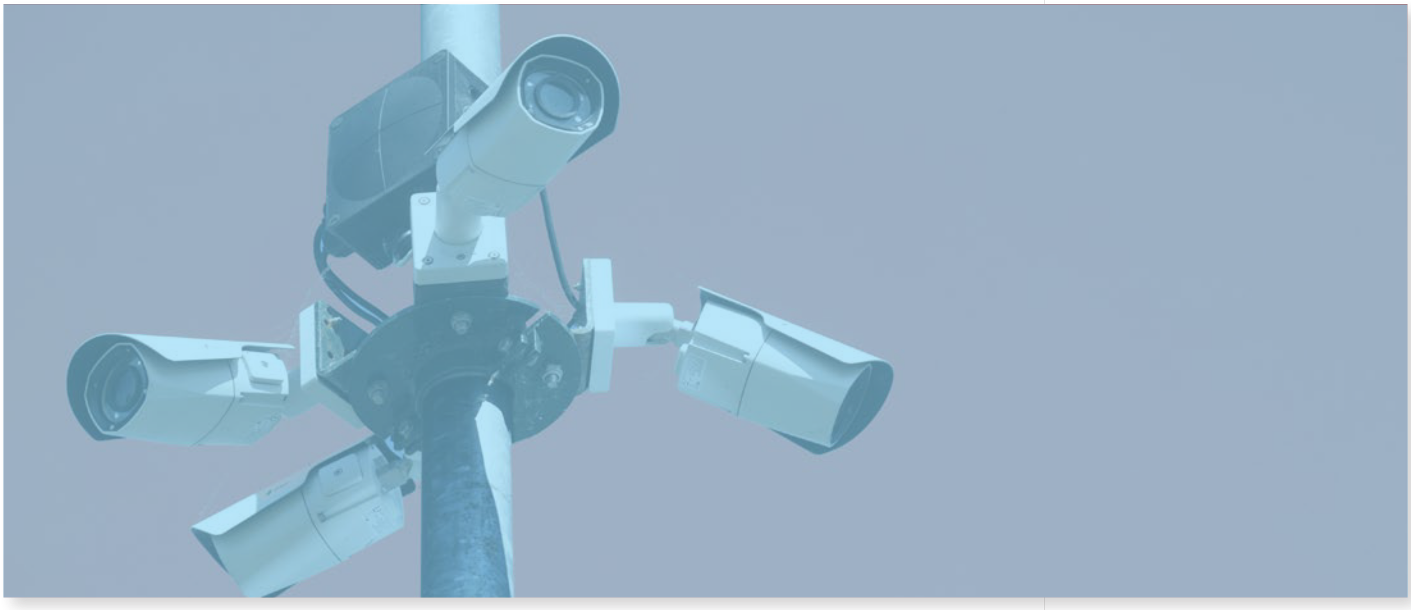
Policymakers can provide value by developing baseline requirements for safeguard and oversight functions, and ensuring that they are acted upon throughout each individual area of consideration.

TIPS FOR THE POLICYMAKERS

Documentation is important for policymakers, to ensure not only that system operators are taking the necessary actions, but also to ensure proper policy deliberations. Discussions and recommendations that are not based on properly documented evidence have a high likelihood of errors and are generally unreliable for policy consideration.

APPENDIX

This appendix provides additional depth on the areas of activity/areas of common concern framework where additional face recognition-specific insights may be helpful.



ADDITIONAL DATA-RELATED CONCERNS

RESEARCH AND DEVELOPMENT

One of the roles of face recognition research is to observe the performance of the technology on hard problems while insulating society from the potential side effects of using the technology. Each problem or scientific question requires specific data to enable the research. Scientists must therefore collect face images under controlled conditions to support their research.

In the early days of face recognition, researchers would manually collect these images from volunteers. They would design collection protocols and human subject consent forms, which would need to be analyzed and approved by the researcher's IRB[23] prior to collection. Most IRBs provide a common experimental protocol template where researchers document all details associated with the experiment, such as the purpose and benefits of the collection, the size and composition of the experimental group, risks to subjects, and data protection. The researcher's submission goes through a comprehensive evaluation by experts in the research field and human experimentation. If the benefits of the research outweigh the risks, the protocol is then approved.

As error rates diminished over time, researchers needed vastly larger data sets and had to shift their focus to leveraging existing data sets, often gathered through operational activities. One of the challenges presented by this need is that data use consent shifted from the individual to the owners of the data sets, with the use still analyzed and governed under the researcher's IRB oversight.

A recent trend in face recognition research has been the proliferation of AI, which requires vast amounts of data to train the recognition algorithm. AI has dramatically lowered the error rates within commercially provided products and has also enabled the creation of additional algorithms by those new to the field. One of the challenges presented by this trend is that some of these new researchers are coming from a purely AI background and are not as mature in handling PII as previous researchers – or possibly do not have IRB capabilities to govern and oversee their work.

Another recent trend in face recognition research has been understanding algorithms' performance differentials across multiple demographic groups. Face recognition research geared toward understanding performance differentials for different demographic groups should seek to understand the source(s) of these differentials, as well as identify research and/or operational approaches to mitigate or eliminate these differentials – just as has been done for other performance differentials (such as pose, lighting, or aging). The challenges to researchers in this area include a lack of necessary data and difficulties in overcoming that data gap via new collections. Policymaker assistance in identifying potential repositories and making them available would greatly enhance the community's ability to minimize these measured differentials.

TESTING

All issues discussed in Research and Development above apply to this area as well, but it is important to note that data used for testing must be different than that used for research, or else results will be skewed.

Data selection for testing purposes is not only complicated, but critical. Incorrectly selected data can easily skew results significantly, rendering the results of the test unusable. Data sets can be improperly manipulated to either boost or discredit performance of a face recognition algorithm. Therefore, special attention should be placed on the source, content, and size of the data set. Any type of data manipulation must be documented.

TIPS FOR THE POLICYMAKERS

Ensuring adequate, consistently used protocols for collecting and using biometric data in research is in need of focus by policymakers.

[23] IRBs are composed of researchers, subject matter experts, and community members dedicated to the review and approval of proposed human subject experiments. Technical or research publishers (such as Nature, Institute of Electrical and Electronics Engineers, and Science) do not publish work involving human subjects unless the work was approved by an IRB. All research universities and major federal agencies, such as the Department of Defense, Department of Energy, and the National Institutes of Health, have their own IRBs. Industry has access to private IRBs

The amount of data used directly affects how precise the test results will be. More data enables more trials, which produce more precise and statistically significant results. An evaluation's error measure is as important as the average performance because it sets expected performance boundaries. Small data sets tend to provide results with large error measures, because there is less confidence with fewer samples.

Data for technology evaluations should enable targeted analysis on a variety of factors, such as variations in pose, illumination, expression, and occlusions; face images of people of different sex, age, and race; of people from different geographic regions; captured at different distances; of a person at different ages (temporal data); and captured under different conditions. Results from technology evaluations should primarily be used to understand where future research needs to be prioritized or to indicate which algorithms to select for next-stage scenario evaluations. Claims that equate a technology evaluation result to any eventual operational accuracy expectations will often be misleading.

Data used for scenario evaluations must be like what would be expected in operations for the results to provide insight on how the system will perform. This includes representation from anticipated demographic users as well as anticipated operational considerations (such as bad poses or lighting).

INITIAL IMPLEMENTATION CONSIDERATIONS

At this stage, meaningful deliberations on the data necessary for operations begin. These deliberations include several topics:

- What types of data input will be operationally available?
  - Will operators be able to ensure that input images are ideal, or will they have to work with whatever unconstrained image they can get?
- What data standards should be used?
- What data will be held within the system?
  - Who will be represented in the database? Are the subjects only U.S. citizens? Will any be minors?
  - What is the sensitivity of this data?
  - Who owns the data? (Could be multiple)
  - Will identities be linked to the biometric samples within the same database?
- What other identifying information will be collected alongside the face data?
- Where will the data be stored and secured?
  - Will the database be in a secure, access-controlled facility?
  - Does the entity have the requisite skills and experience to manage this data?
- Who will have access to the data?
- Are all accesses to the database logged?
- Is the point of collection trusted?
- How will the integrity of the data be confirmed?
- Has the data in the system been specifically authorized by the responsible organization for use in that database and approved for this use by their IRB?
- Will the data in the system provide timely and accurate returns or answers to users of the system (e.g., if they expect criminals in the system, is this what the system provides?)?

- How similar are the characteristics of the operational input data to that previously enrolled into the system? How will variances impact performance?
- Was the algorithm tested on operationally relevant data and verified to perform functionally within the requirements to meet the needs of the system?

Additional Study Guide

Additional Considerations for Use Case Category 1: Verifying Claimed Identities.

Depending on the design of a system, the data held may come down to a single image or set of images seeking verification of a single individual e.g., personal identity verification cards, passports). In these scenarios, the risks of data being captured in invalid formats or being operationally irrelevant are minimal, as the impact will only yield denials that can be individually remedied via subsequent access attempts or alternative authentication mechanisms.

System operators will also need to determine if individual transaction data really needs to be recorded, and if so, how long it must be kept.

Additional Considerations for Use Case Category 2: Identity Determinations.

Investigative data is often comprised of large varieties of data, ranging from mug shots and passport photos to completely unconstrained photos, depending on the nature of acquisition. Ensuring that systems are resilient to these types of data and/or that the processes and system design accommodate the variances in data is paramount to the successful, and correct, utilization of face recognition technologies.

Additional Considerations:

- Is this investigative search of the data appropriate or authorized?
- How will the data be accessed? What are the appropriate capabilities and limitations within that process?
- What records of the search need to be kept by all involved parties?

Additional Considerations for Use Case Category 3: Watchlist.

- What are the retention needs and policies for the original image?
  - How does this vary depending on output of the resultant biometric comparison?
  - How should original imagery that includes multiple individuals, or of varying levels of concern, be handled?

DISPOSAL

One of the most important policy considerations when terminating a face recognition system is deciding what to do with the operational data. The answers will vary considerably based on the individual application and the makeup of its data. Key considerations include:

- Under what conditions/restrictions was the data collected?
- Can the data be separated into individual files, or has it lost its provenance?
- Will the data be transferred into a new system?

- Do system operators need to go back to the subjects of the system to inform them of the disposition of their data?
- Who owns the data?
- Has data from this system been shared with other systems, and what (if anything) do system operators need to do about that sharing?
- Can the data be completely deleted?
- How can the agency verify that unnecessary data is deleted?

Additional Study Guide

Additional Considerations for Use Case Category 1: Verifying Claimed Identities.

Data from this use case can have value that lasts past the termination of the system. For example, future security investigators could require access to logs of requests and authorizations.

Additional Considerations for Use Case Category 2: Identity Determinations.

In the investigation use case, data is often transitioned to a case management system and retained according to the rules governing the specific investigation.

Additional Considerations for Use Case Category 3: Watchlist.

Most data collected in this use case is retained to the greatest extent possible to enable future analyses when currently unknown links are later discovered. Policymakers need to ensure that this capability is available while also adding restrictions to curtail unnecessary use.

ADDITIONAL PRIVACY-RELATED CONCERNS

RESEARCH AND DEVELOPMENT

There are no laws specific to face recognition research and privacy considerations, although general-purpose best practices and guidance exist.[24] Federally funded face recognition research must also meet regulations for managing PII and conducting human experimentations, which is a high standard.

Additional Policy-Level Considerations Include:

- Face recognition research geared toward the refinement of any algorithm or system for a specific operational purpose needs to be predominantly based on data samples that reflect the demographics of that application's user community. Both working to identify this need and obtaining the necessary data to overcome it can introduce privacy concerns.
- When performance differentials are identified for different demographic groups in research, testing, or operations, they should be documented and steps should be taken to understand the source(s) of these differentials, as well as to identify research and/or operational approaches to mitigate or eliminate them.
  - Algorithms alone may not be able to mitigate, so training operators to understand these impacts is necessary.

[24] See discussion in the Privacy and Civil Liberties section

- Any biometric research requires the collection of ground-truth data (i.e., multiple samples from the same individual must be collected and retained in a manner that protects that individual's privacy and civil liberties).
  - Approval of data collection through IRBs is an important mechanism to ensure that researchers protect subjects' privacy and civil liberties.
  - Subjects who volunteer to participate in research should grant consent to have their data collected as a means of ensuring their privacy and civil liberties.
  - For non-volunteer situations, care should be taken to ensure that the collection of biometric samples (such as mug shots or border crossing data) have been specifically authorized by the responsible party for use in research. Maintaining security and access restrictions is of utmost importance.
  - Data collected for biometric research should be anonymized to the extent possible. Demographic metadata such as ancestry and age cannot be disconnected from the biometric samples if the data is to be binned/examined based on that metadata, but other PII (such as names and addresses) should be replaced with non-identifiable markers.
- Personnel with access to biometric research data should be trained in conducting human subject research, including guidelines and procedures for protecting privacy and civil liberties.

DISPOSAL

The privacy and civil liberties risks of face recognition systems do not end with the termination of the system. In several cases where operational data and results must be stored, they have been transitioned into security or other investigative files whose retention requirements can outlive the face recognition system itself. Continuing risks from this information include:

- Breaches of data. Even though this would be a breach of security or investigative files rather than the face recognition system itself, it may not be highlighted as such.
- Misuse of the data. Although this risk is lower since the data is in storage rather than current use, a concern remains that this data could be used in a manner outside the face recognition system's authorization or policies.

There are no laws or federal policies on privacy or civil liberties concerns specific to face recognition systems that are being terminated. However, in the case of federally managed systems, there are rules and best practices (written for all federal systems) that are explained in the Federal Acquisition Regulations and the Defense Acquisition Guidebook.

Questions for policymakers to consider include:

- Is the project completely closing or being replaced by another program?
- When individuals began be entered into the system, what promises were made about the use of their PII after termination of the project?
- When data is to be deleted, what methods will be used and what oversight is planned to ensure that deletion is performed properly?

Additional Study Guide

Additional Considerations for Use Case Category 2: Identity Determinations.

In this use case, results (and possibly data) from the face recognition system have almost always been stored within investigative files and possibly shared with other jurisdictions collaborating on the investigation. Security and misuse concerns will continue beyond system termination.

Additional Considerations for Use Case Category 3: Watchlist.

Data will need to be preserved and deleted according to the rules under which it was collected and with consideration of anticipated future needs.

ADDITIONAL SAFEGUARDS AND OVERSIGHT-RELATED CONCERNS RESEARCH AND DEVELOPMENT (AND TESTING)

IRBs that enforce each institution's policies provide the predominant safeguard and oversight for face recognition research and testing. Most IRBs impose an arduous approval process. Most institutions require team members to receive certification for human subject experimentation. The most common certification is offered by the Collaborative Institutional Training Initiative. The depth and breadth of the training varies by institution. The investigators submit to the IRB what is known as the experimental protocol, which includes the objectives, members of the research team, description of the population pool and size, risks, procedures, instrument and materials involved, recruiting material and strategy, compensation to subjects, procedure for consent, data protection and storage, benefits, and many other items.

All methods, decisions, and assumptions need to be supported by previous scientific work or known best practices. The IRB has the power to require reviews from the institution's local biosafety or health and safety officers. After a comprehensive review of the application and support documents, the IRB approves the proposed research or testing if the benefits exceed the risks and subjects' privacy is preserved – unless otherwise clearly noted during consent. Although not all experiments must end in a direct benefit to the participants, there must be a clear benefit to society and the nation for the research to be approved.

IRBs can approve simple research or testing protocols with minimum risks within weeks, while experiments that can have a small impact on the subjects' health, safety, and privacy can take several iterations and months for full approval. The IRB is critical to ensure fair and safe treatment of subjects. Moreover, discussions between the IRB and the investigators result in stronger and better conceived experiments.

As IRBs are internally managed and generally focused, and there are no legal requirements that they must exist for face recognition research specifically. This is a potential area for legislator or policymaker attention. Promulgating consistent requirements or best practices, as well as crafting guidance to help ensure that IRBs can perform this duty properly for face recognition research, would be a positive step.

INITIAL IMPLEMENTATION CONSIDERATIONS

This is the most important stage for safeguards and oversight of an operational system, as agencies must determine how they will be implemented so that they can be designed into the system architecture and processes on day one.

Plans for the proper safeguards and oversight of any face recognition system in an agency's initial steps toward operations should be carefully derived and put into place, as the downstream impacts to system design, operations, and maintenance will be paramount. Getting ahead of the curve in defining these safeguards and oversight provisions will provide transparency and/or accountability to the agency. For example, if these safeguards and oversight mechanisms are designed properly, the agency will be able to do the following, while also providing reassurance to the public and lawmakers:

- Ensure that the correct organizations can store the data for authorized use cases.
- Ensure that only authorized users have access to the systems and data that are central to their mission.
- Ensure that results are disseminated only to authorized parties.
- Ensure that oversight is established to vet use cases, data, and security of face recognition systems.

Policies, laws, and other regulations for safeguarding and oversight vary greatly from agency to agency. For example, federal law enforcement agencies must operate within stringent requirements that are defined across systems and within organizations, whereas individual commercial organizations may have more flexible and varied requirements, depending on their organizational structure, size, and intended application.

Existing standards and best practices for safeguards and oversight are often managed across individual organizations and agencies, due to the various applications, data, and missions they are acting upon. The National Telecommunications and Information Administration[25] and Government Accountability Office[26] have released guidance for commercial use of face recognition. The private sector has also attempted to establish oversight recommendations on its own through several publications, including:

- *Ethical Principles for Biometrics[27] and Good Practice Framework[28] from the Biometrics Institute*
- *Privacy Principles Whitepaper from the FIDO Alliance[29]*
- *Facial Recognition Policy Principles from the U.S. Chamber of Commerce[30]*
- *Principles for the Responsible and Effective Use of Facial Recognition Technology from the Security Industry Association[31]*

Additional considerations for policymaker attention include:

- What is the appropriate level of safeguards and oversight for different use cases and levels of data sensitivity?
- How is the system itself protected from unauthorized utilization? Include spaces such as cyber attacks, user authentication and security, and vetted use cases.
- How does an agency vet what techniques and systems individual users utilize for their use cases (i.e., to exclude rogue users)?
- What vetting mechanisms does an agency have to determine authorized use, and to report unauthorized uses of the system? Will uses of the system be logged for future analysis?

[25] *Privacy Best Practice Recommendations For Commercial Facial Recognition Use.* 2016. National Telecommunications and Information Administration, https://www.ntia.doc.gov/files/ntia/publications/privacy_best_practices_recommendations_for_commercial_use_of_facial_recogntion.pdf.

[26] *Facial Recognition Technology: Privacy and Accuracy Issues Related to Commercial Uses*. 2020. U.S. Government Accountability Office, https://www.gao.gov/assets/710/708045.pdf

[27] *Ethical Principles for Biometrics*. 2019. Biometrics Institute, https://www.biometricsinstitute.org/ethical-principles-for-biometrics/. Accessed October 29, 2020.

[28] *Biometrics Institute Good Practice Framework*. 2020. Biometrics Institute, https://www.biometricsinstitute.org/biometrics-institute-good-practice-framework/ Accessed October 29, 2020.

[29] *Privacy Principles*. 2014. Fast Identity Online Alliance, https://fidoalliance.org/wp-content/uploads/2014/12/FIDO_Alliance_Whitepaper_Privacy_Principles.pdf

[30] *Facial Recognition Policy Principles*. 2019. U.S. Chamber of Commerce, https://www.uschamber.com/sites/default/files/ctec_facial_recognition_policy_principles_002.pdf.

[31] *Principles for the Responsible and Effective Use of Facial Recognition Technology*. 2020. Security Industry Association, https://www.securityindustry.org/wp-content/uploads/2020/08/SIA-Principles-Responsible-Ethical-Facial-Recognition-Usage.pdf.

- What technical assessments were leveraged to determine the proper utilization of specific technologies within the system? Are these re-evaluated periodically to ensure that they meet requirements and intended use of the system?
- When does internal oversight become insufficient, thus requiring outside or independent oversight?
- How frequently must the safeguards be checked to ensure proper operation? How and how often will the oversight be performed?

Additional Study Guide

Additional Considerations for Use Case Category 3: Watchlist.

One important oversight consideration is the criteria for how subjects are added, and in what cases subjects are removed from the watchlist, due to the negative outcomes of matches to a subject on the watchlist.

DISPOSAL

Safeguards and oversight of termination activities remain important, though they are often overlooked due to lack of interest or priority. While there are no common requirements specific to terminating the safeguards and oversight of face recognition, there are general best practices that can be leveraged.

As mentioned in prior discussions, data and results from the face recognition system are often retained, and occasionally used, beyond the lifecycle of the face recognition system itself. Policymaker considerations include:

- Deciding what notifications need to be provided about the termination, and to whom
- Reassessing data retention and sharing decisions
- Ensuring that formal responsibility for data ownership is transitioned
- Ensuring that documentation remains available for future investigative and trial purposes
- Deciding which of the existing safeguards and oversight responsibilities and processes change at termination, and what new ones must be established

ADDITIONAL REPORTING CONSIDERATIONS

RESEARCH AND DEVELOPMENT

The hallmark of quality research is the ability of others to repeat the experiments and achieve the same results, thus enabling researchers to continuously build upon each other's work. This requires a great deal of documentation, which is usually published in technical journals and conferences. Much, but not all, published work first undergoes peer review. If the work is sound and exceeds certain novelty expectations, it is recommended for publication to the editor. Not all publications have the same stature, however. Those that have historically had the most impact receive the most potential papers, and thus have the most difficult peer reviews. Papers found in these publications are certain to be both credible and impactful.

It is also important to note that research that is not published in top journals, or anywhere for that matter, is not automatically discreditable. The ease of publishing online, combined with the desire to make impacts faster and to control the rights to their own research, has pushed more and more researchers to self-publish. There are also executive branch-led efforts to reassess the nation's research publishing paradigm.

TIPS FOR THE POLICYMAKERS

Not all research is equal.
Good research will be found, championed, and leveraged by other researchers over time.
If research produces a novel outcome and no one else in the community has begun using it (or if they are actively criticizing it), then that indicates that something is amiss, and those results should be discounted.

TESTING

Test results that lack sufficiently detailed reports cannot be trusted. Just as in research reporting, test reports must contain enough detail that any other learned professional could recreate the test and obtain the same results. While there are some common considerations, each type of evaluation also requires specialized reporting:

- Common considerations:
  - What was the makeup of the data, and was it sufficiently detailed that it could be recreated?
  - What steps were taken to ensure that the data selected was appropriate for the experiment(s) and did not introduce issues that could produce inaccurate results?
  - Which algorithms, and which versions of the algorithms, were used?
- Technology evaluations:
  - How was the data parsed for use in each experiment? Was enough data used to make the results statistically significant?
  - How much information was given to providers in advance, so that they could tune their algorithms for this specific experiment?
- Scenario evaluations:
  - What was the physical makeup of the test environment?
  - What were the tactics, techniques, and procedures (TTPs) for the experiment?
  - What was the demographic makeup of the test subjects? What training and/or prior experience did they have?
  - Given the size of the database and number of participants, what is the confidence factor in the measured results?
- Operational evaluation:
  - What was the physical makeup of the operational environment?
  - What were the TTPs for the experiment?
  - If specific test subjects were used, how were they integrated into the general user population, and how does that affect results?

Answers to these questions will help policymakers assess how credible each test result may be, as well as determine how the results should be interpreted.

TIPS FOR THE POLICYMAKERS

Policymakers should be skeptical of any test result when the experimental population size is small, when an algorithm is tested outside its intended design, or when performance is not accompanied with measures of potential error (e.g., the error rate for system A is 12% ± 1% accurate).

INITIAL IMPLEMENTATION CONSIDERATIONS

Work performed during the initial implementation considerations stage sets the requirements and expectations for system design and operation (including privacy and civil liberties protections). Merely considering and verbalizing these requirements and expectations is not sufficient, however. They must be put into writing to be effective, and the more specific the better.

Establishing regular reporting protocols that address the challenges of their mission and concerns of privacy and civil liberties groups will ensure the proper execution of face recognition systems; weak reporting and accountability often causes startling issues for face recognition systems. Additionally, if reporting is not considered as part of the initial operation of the system, retrofitting to add such reporting can be costly. It is imperative that reporting mechanisms consider the applications and needs of agencies and their use of face recognition technology and provide comprehensive and accountable, but reasonable and responsible, reporting requirements for systems.

Additional considerations for initial implementation considerations documenting include:

- What legal authorities permit this operation (if required)? What policies regulate its operation?
- How will the system operate (e.g., process flow, accessing and storing data, making decisions)?
- What needs to be designed in to enable needed safeguards and oversight?
- When standing up an operational system, what aspects of reporting need to be targeted for initial deployment?
- What ongoing reporting mechanisms must be accommodated on a regular basis (annually, monthly, etc.)?
- Who is the target audience and managing entity for any reports created?

Additional Study Guide

Additional Considerations for Use Case Category 1: Verifying Claimed Identities.

- What user training will be required?

Additional Considerations for Use Case Category 2: Identity Determinations.

- What processes should be required to request, review, and approve inquiries? How will that vary based on the type of investigation?
- How will face recognition results be conveyed?
- Who is responsible for storing information about each request and its result?

Additional Considerations for Use Case Category 3: Watchlist.

- What are the objectives and limitations of this individual use of face recognition, based on operational, security, privacy, and civil liberties considerations?
- How should hits be handled? Documented?
- What information from this use should be destroyed or kept (and for how long)?
- What is the expectation for post-use review?