JUNE 2021

How Does Facial Recognition Work?

A Primer

By William Crumpler and James A. Lewis

Executive Summary

1. **What Is Facial Recognition?** Facial recognition is a way of using software to determine the similarity between two face images in order to evaluate a claim. The technology is used for a variety of purposes, from signing a user into their phone to searching for a particular person in a database of photos.

2. **What Is Facial Characterization?** Facial characterization refers to the practice of using software to classify a single face according to its gender, age, emotion, or other characteristics. Facial classification is distinct from facial recognition, whose purpose is instead to compare two different faces. Facial characterization is often confused with facial recognition in popular reporting, but they are actually distinct technologies. Many claims about the dangers of facial recognition are actually talking about characterization.

3. **How Does Facial Recognition Work?** Facial recognition uses computer-generated filters to transform face images into numerical expressions that can be compared to determine their similarity. These filters are usually generated by using deep "learning," which uses artificial neural networks to process data.

4. **How Accurate Is Facial Recognition?** Facial recognition is improving rapidly, but while algorithms can achieve very high performance in controlled settings, many systems have lower performance when deployed in the real world. Summarizing the accuracy of a facial recognition system is difficult, however, as there is no single measure that provides a complete picture of performance.

5. **What Are Similarity Scores?** Similarity scores provide feedback to human operators about how similar the algorithm believes two images are. These scores can be misunderstood and are often treated as providing more authoritative information than they really do because of something known as the "prosecutor's fallacy."

6. **What Are Comparison Thresholds?** Facial recognition systems face a trade-off between low false negative rates and low false positive rates. Comparison thresholds are a way of using the similarity scores calculated by facial recognition algorithms to tune a system's sensitivity to these two types of errors. Thresholds are adjusted to account for trade-offs between accuracy and risk when returning results to human adjudicators.

7. **Is Facial Recognition Biased?** Demographic differences in facial recognition accuracy rates have been well documented, but the evidence suggests that this problem can be addressed if sufficient attention is paid to improving both the training process for algorithms and the quality of captured images.

8. **What Does This Mean?** Facial recognition is usually discussed only in the context of its most dystopic applications, but it is a multifaceted tool that can be applied to a range of different problems. Facial recognition is used to aid human decisionmaking rather than replace it. Human oversight helps to mitigate the risk of errors. Operators need to understand how system performance can be affected by deployment conditions in order to put in place the right safeguards to manage trade-offs between accuracy and risk. A better understanding of the issues covered in this report will help ensure this technology can be deployed safely in ways that let us capture its benefits while managing risks.

Introduction

As an increasing number of organizations begin to use facial recognition technologies (FRTs), concerns have mounted over the potential risks the technology may pose to privacy and other civil liberties. At all levels of government, policymakers have begun to propose new rules and regulations to govern the use of FRTs and manage any risks. It is important that these efforts be grounded in fact about how the technology works, since much public discussion so far has been muddled by exaggerations about the technology's performance, misunderstandings about the details of its operation, and conflation of different types of systems. This paper provides an examination of how the technology works and how to understand questions about its performance and operation.

What Is Facial Recognition?

Facial recognition is a subfield of computer vision research focused on building software systems that can analyze the similarity between faces in images and video. In practice, facial recognition tools can be thought of as a way to evaluate a claim involving a particular person. Those claims can be anything from "is this person who they say they are?" to "is this person contained within this database?" or even simply "has this person ever been seen by the system before?" While this can sometimes be done in an entirely automated manner, facial recognition is usually deployed in combination with human examiners who are responsible for reviewing and adjudicating the decisions returned by the software.

Today, almost all new facial recognition systems are built with the help of deep learning, a form of machine learning that uses artificial neural networks to process data. Facial recognition developers use deep learning to create software programs capable of transforming face images into numerical expressions that can be compared to determine their similarity. Facial recognition is a form of biometric identification, but it is important to note that not all biometric processing involves the use of deep learning.

Importantly, facial recognition is different from facial *characterization* (also sometimes referred to as facial analysis). In facial recognition, algorithms are used to compare the similarity of two faces. In facial characterization, algorithms are used to classify a single face according to its gender, age, emotion, or other characteristics. Used on its own, facial characterization can be used to anonymously profile individuals for purposes such as counting the number of men and women entering a particular store or providing data about how different demographic groups respond to a product or advertisement.

Facial characterization is a distinct technology with its own separate development process, uses, and risks, but people sometimes use the terms facial characterization and facial recognition interchangeably when they are in fact very different. There are important conversations to be had about how to govern characterization systems—especially given recent attempts to use it for highly questionable purposes such as classifying people by **ethnicity** or "detecting" an individual's **sexual orientation, political orientation,** or **criminality**—but these should be undertaken with an awareness that facial characterization and facial recognition are separate technologies.

One of the most common uses of FRT is *verification* (also known as 1:1 matching), where the technology is used to confirm whether a person is connected to a specific identity record. Examples of verification are when a person uses their face to unlock their smartphone, sign in to a banking app, or verify their identity when passing through airport security. When a person logs in, the system takes a picture of their face and then compares it with the image on record for that person. If the two faces match, the person is then granted access. Comparison photos are usually either taken when a person first signs up for the service or drawn from a trusted source such as a passport or national identity registry.

Identification (also known as 1:N or 1:many matching) is when facial recognition is used to determine whether a record for an unknown individual exists in a larger database of known faces. The most well-known example of identification is the police practice of using facial recognition to generate a lineup of potential suspects based on images or footage of a crime. However, law enforcement has also used identification to search for missing persons, identify deceased individuals, and de-duplicate database records. Identification can also be used by the private sector to enforce blacklists (such as when a casino monitors its customers to detect gambling addicts) or whitelists (such as when a building's management wants to automate the process of granting access to employees or residents). Identification almost always incorporates human review, either by requiring a trained human operator to select a match from a list of options presented by the software or by allowing individuals to appeal decisions they disagree with to human adjudicators.

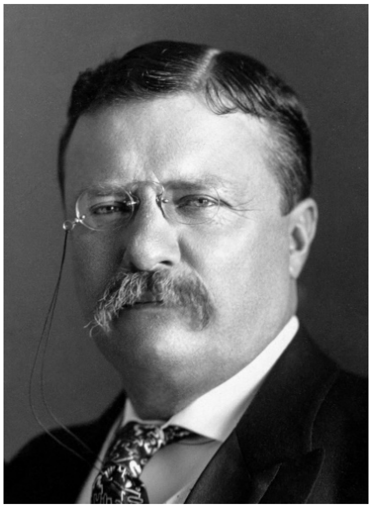
Identification does not necessarily supply any information about who a person is. That is, FRT can be used in ways that do not actually involve collecting or linking any personal data about the person in question. For example, some retail stores may create customer tracking systems that allow the store to recognize return customers and track in-store behavior for marketing analysis but not link that data to any biographic information such as a name, address, or purchasing history. The only thing the store facial recognition system would recognize is that visitor #12345 had returned on a certain day. It would not have any way to tell that visitor #12345 was Jane Smith of 678 Main Street. Similar systems can also be used in the opposite way—to identify when a person has *not* been seen before. An example of this is the Beijing **park** that was experiencing issues with visitors taking too much toilet paper and decided to install toilet paper dispensers with FRT that first check to ensure that a person had not been encountered by the system in the past nine minutes.

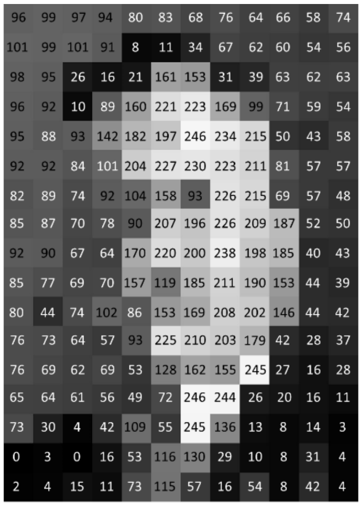
How Does Facial Recognition Work?

Facial recognition works by transforming an image of a face into a numerical expression called a *template* that can be used to compare the similarity of face images. By comparing the templates of different faces, it is possible to determine whether two given faces belong to the same subject, similar to how one might compare fingerprint records. The process of generating these templates is often described as a matter of locating certain key points on a face and measuring the distances between them, but this is not accurate.

While some older attempts at building facial recognition systems did take this approach, today's systems are far more sophisticated. Modern facial recognition programs instead generate face templates by passing an image through a series of computer-generated "filters."

To understand how this works, it is important to first remember that a computer does not interpret a picture of a face the same way a human would. A computer does not see a face as a collection of colors, shapes, and features, but rather as a matrix of numbers. These numbers are arranged so that each one describes the lightness or darkness of a certain pixel. The goal of facial recognition is to find a way to reliably recognize a face from the way these numbers are organized within the matrix.

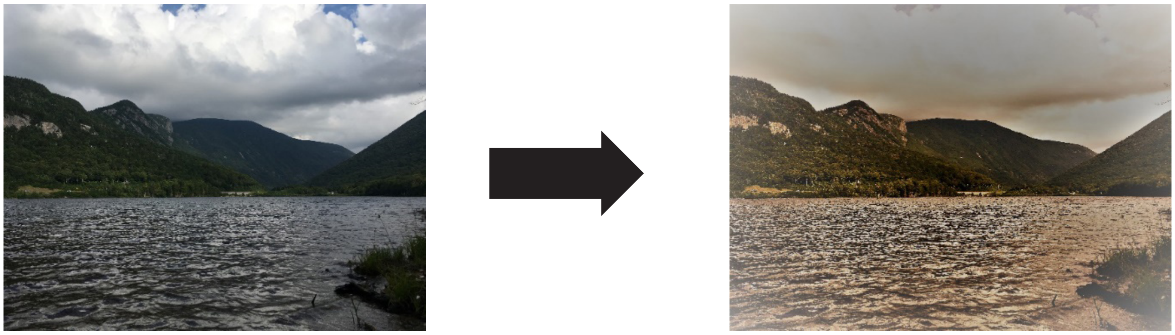




96	99	97	94	80	83	68	76	64	66	58	74
101	99	101	91	8	11	34	67	62	60	54	56
98	95	26	16	21	161	153	31	39	63	62	63
96	92	10	89	160	221	223	169	99	71	59	54
95	88	93	142	182	197	246	234	215	50	43	58
92	92	84	101	204	227	230	223	211	81	57	57
82	89	74	92	104	158	93	226	215	69	57	48
85	87	70	78	90	207	196	226	209	187	52	50
92	90	67	64	170	220	200	238	198	185	40	43
85	77	69	70	157	119	185	211	190	153	44	39
80	44	74	102	86	153	169	208	202	146	44	42
76	73	64	57	93	225	210	203	179	42	28	37
76	69	62	69	53	128	162	155	245	27	16	28
65	64	61	56	49	72	246	244	26	20	16	11
73	30	4	42	109	55	245	136	13	8	14	3
0	3	0	16	53	116	130	29	10	8	31	4
2	4	15	11	73	115	57	16	54	8	42	4

96	99	97	94	80	83	68	76	64	66	58	74
101	99	101	91	8	11	34	67	62	60	54	56
98	95	26	16	21	161	153	31	39	63	62	63
96	92	10	89	160	221	223	169	99	71	59	54
95	88	93	142	182	197	246	234	215	50	43	58
92	92	84	101	204	227	230	223	211	81	57	57
82	89	74	92	104	158	93	226	215	69	57	48
85	87	70	78	90	207	196	226	209	187	52	50
92	90	67	64	170	220	200	238	198	185	40	43
85	77	69	70	157	119	185	211	190	153	44	39
80	44	74	102	86	153	169	208	202	146	44	42
76	73	64	57	93	225	210	203	179	42	28	37
76	69	62	69	53	128	162	155	245	27	16	28
65	64	61	56	49	72	246	244	26	20	16	11
73	30	4	42	109	55	245	136	13	8	14	3
0	3	0	16	53	116	130	29	10	8	31	4
2	4	15	11	73	115	57	16	54	8	42	4

One way of achieving this is by using filters. Today, the idea of image filters is most familiar in the context of smartphone camera apps. These kinds of filters can take an image and adjust its color, contrast, or other visual details by going through the matrix of numbers that make up an image and slightly tweaking them all according to a pre-determined set of rules that create a particular visual style. The result of this manipulation is that when you put all the numbers back together, the image looks slightly different to our eyes.



The filters used for facial recognition are based on the same idea of manipulating pixel values according to a set of programmed rules. However, they differ in that their purpose is not to tweak the visual details of a face image but rather to transform the image into a simplified "fingerprint" that distinctively represents that face. To understand what this means, consider the case of the following sets of numbers:

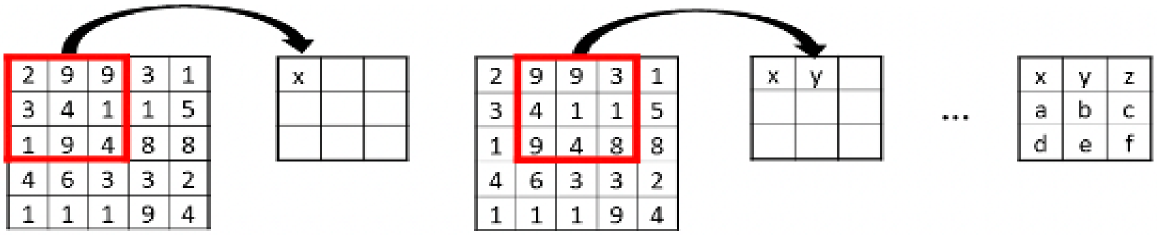
List 1: [8, 1, 2, 4, 2, 5, 6, 7, 8, 7, 8, 9, 8, 4, 2, 1, 1, 6, 6, 2, 2, 2, 1, 6]

List 2: [6, 8, 8, 3, 4, 1, 7, 8, 3, 4, 1, 7, 4, 2, 5, 8, 2, 3, 3, 2, 7, 7, 9, 8]

Imagine you were given the task of comparing these two lists to determine whether they contained the same numbers (just in a different order). How would you go about it? One answer may be to manually go through each list to check off each number against the opposite list, but this would take a very long time and would not scale well if you were asked to do many thousands of these comparisons. A better solution would be to find some way of simplifying each list into a form that was easier to compare immediately. One option might be to use the average of the values in the lists. If the two lists contained the same numbers, their averages should also be the same. You could calculate that List 1 has an average value of 4.5, whereas List 2 has an average value of 5. By using the average value of the list as a kind of fingerprint, it is possible to quickly and easily determine that the lists contain different sets of numbers.

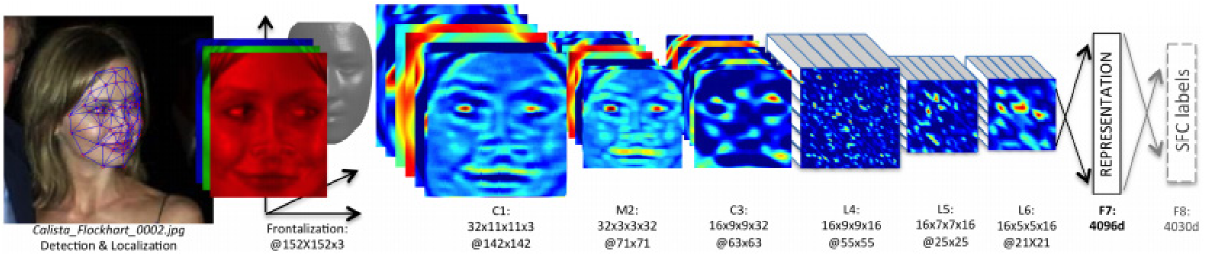
Of course, the average of a list of numbers is not a unique representation of that list. Even if you were presented with two lists that both had an average value of five, you would not be able to know for certain that they contained the exact same numbers. One list may hold 20 fives while the other held 10 ones and 10 nines. A better operation would be able to reduce those lists into a representation that was more distinctive, and which would not be mistaken for anything other than a truly matching list of numbers. This is the mission of facial recognition developers: to create a way of transforming the information contained in a digital face image into a simplified but still highly distinctive representation. These templates would ideally be unique in the sense that no two faces should ever produce highly similar templates, and they should be robust in the sense that different instances and angles of the same face should always lead to highly similar templates. Currently, the templates created by modern FRT systems are neither unique nor fully robust but getting closer to these goals is the focus of current research efforts.

The operation performed by a facial recognition system to produce a template from a digital image is far more complex than simply taking the average of all the numbers in the matrix of pixel values. The template generation process in modern facial recognition systems involves applying a series of filters that move across the image and perform calculations on pixel values to yield simplified representations, as illustrated below.



In the illustration above, the numbers shown represent the pixel values of the facial image. The red box is the "filter," containing numbers or a computational process not illustrated. The output values (x, y, z, and so on) are the result of multiplication and additions or some other operations (such as "find the largest number") on the image pixels within the red box.

By repeating this process over and over on the image using multiple different filters, facial recognition models are able to eventually produce a simplified template that is easily comparable while still being distinctive to the subject and hopefully robust to the quality of the images. An example of an algorithm that works this way is Facebook's DeepFace system, illustrated below by Facebook in a **recent report**, which puts images through seven stages of processing and begins by applying 32 different filters to three different versions (one version for each primary color) of the target image. The complexity of this approach shows just how far this technology has come from early attempts to simply measure the distances between key points on subjects' faces.



Source: Yaniv Taigman et al., "Deepface: Closing the Gap to Human-Level Performance in Face Verification," Conference on Computer Vision and Pattern Recognition (CVPR), Facebook, June 24, 2014, https://research.fb.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/.

The key question driving this process is what operation the filters perform on the pixel values to create the simplified representations. Simply averaging the values inside the filter as it scans across the image will not work. In the early days of facial recognition research, scientists were responsible for manually selecting the filters that would be applied. Today, however, these filters are instead chosen by the computer itself.

Modern facial recognition developers use deep learning to automate a process of trial and error that helps identify the best filters for reliably generating robust templates. Training these systems involves providing them with a series of "triplets"—collections of three face images where two of the faces belong to one person and the third belongs to someone else. The system turns each of the three images into a template and then compares their similarity. The system is given the goal of achieving the maximum similarity for the templates coming from the same subject and the minimum similarity for the templates coming from different subjects.

As the system churns through tens or hundreds of thousands of these triplets, the algorithm continuously tweaks the operations its filters perform and then measures whether the changes it made result in better or worse accuracy in correctly determining which of the three images are from the same person. If a change leads to an improvement, the system keeps it, and if performance gets worse, the system will revert to its previous state and then try something else. In this way, the system slowly learns which filters are the best at creating distinctive face templates. At the end of this process, the system arrives at a set of filters that have repeatedly proven their efficacy. At this point, the model is ready to be packaged as part of a software program and sold to operators.

How Accurate Is Facial Recognition?

Facial recognition has improved dramatically in only a few years. Precisely quantifying this progress is difficult, however, because there is no single value that provides a complete picture of a facial recognition system's "accuracy." It is very common for vendors and commentators to use general statistics to describe their facial recognition system's performance, such as claiming that a particular system is "99 percent accurate." But these kinds of assertions do not give a complete picture of the performance operators should expect to see once the system is actually deployed.

For one thing, it is not immediately obvious what a claim like "99 percent accurate" actually means. This is because there are two different kinds of mistakes that a facial recognition system can make: false negatives and false positives. False negatives occur when the system incorrectly says two images of the same person do not match. False positives occur when the system incorrectly says two images from different people are the same person. To take the example of a smartphone that logs you in by scanning your face, an example of a false negative would be when you pick up the phone, and it does not let you in because it thinks you are someone else. A false positive would be when somebody else picks up your phone, and it mistakenly logs them in thinking that they are you.

Therefore, a claim of 99 percent accuracy could mean that out of 100 times the system was presented with a matching face, it only incorrectly rejected one (a false negative rate of 1 percent). Or it could mean that out of 100 imposters who tried to log in to the system, only one ended up being accepted (a false positive rate of 1 percent). It could also mean that out of 100 total judgments (both rejections and approvals), only one was incorrectly decided. This is an important distinction, especially when thinking about high-risk applications such as law enforcement investigations. But even if this information were specified, the number is still not helpful until you also have the context of the tests. Was this 99 percent accuracy measured by testing on high-quality images in a lab or by observing the system's actual operation in a real-world setting?

The most authoritative source of data on the accuracy of facial recognition systems comes from the U.S. National Institute of Standards and Technology (NIST) and their **Face Recognition Vendor Test** (FRVT), a series of evaluations on facial recognition systems that has been ongoing for two decades. Results from the FRVT allow us to gain a sense of how the technology has improved over time. One indicative test in the FRVT involves providing facial recognition algorithms with the photo of a test subject and then asking the system to search through a curated database of 1.6 million high-quality mugshots to return the one the algorithm thinks is most similar. In 2013, the best algorithm returned the wrong photo **4.1 percent** of the time.[1] The leading algorithm, as of March 2021, returned the wrong photo **less than 0.1 percent** of the time.[2] NIST reported that as of 2021, at least 30 developers submitted algorithms that outperformed the leading system in 2013, pointing to broad improvements across the industry. Even when the subject is wearing a mask obscuring more than 70 percent of their face, NIST tests have **found** that leading algorithms can achieve over 97 percent true positive rates, which is equivalent to the performance of state-of-the-art systems on unmasked faces in 2017.

These results are impressive and help explain the recent interest of so many organizations in adopting the technology. However, it is important to note that this degree of accuracy is contingent on a number of factors, most importantly, the algorithm being used, the quality of the images being compared, and the size of the search space. The photos used in the tests described above were taken in dedicated photographic environments with assistance from a human operator. This helped ensure that images had good lighting, that the positioning of subjects' faces was consistent, and that facial features were never unclear or obscured. Facial recognition systems have much better accuracy rates when the images they are comparing are clear

[1] Rank-1 investigation miss rate at N=1.6M mugshot photos for 2013 NEC-30 algorithm (0.041). Table 1.

[2] Rank-1 investigation miss rate at N=1.6M mugshot photos for SenseTime-005 algorithm (0.0009). Table 16.

and consistent. NIST **reported** that when testing instead on images taken from self-operated ATM kiosks where images had a far greater range of quality, error rates often rose to above 20 percent, even for some of the most accurate algorithms.

Importantly, it is not just the images captured during the sign-in attempt which must be high quality, but also the stored target images being used for comparison. For example, if a facial recognition system is set to compare a person against a database of records, the system may experience errors if those records are not regularly updated. Changes in subjects' faces over time can make it difficult to match pictures taken many years apart. NIST's tests on aging in the FRVT have found that median false-negative error rates increase from **4.1 percent** when matching against photos that are 0 to 2 years old to 32.1 percent when matching against photos that are 14 to 18 years old.[3]

Quality effects are important to note when considering how FRTs may perform in different circumstances. In "unconstrained" settings where subjects are unaware their image is being taken, and operators are not able to control the image quality, the accuracy of facial recognition algorithms is much lower than what is measured in a lab. To measure this, in 2017, NIST conducted their **Face in Video Evaluation** (FIVE) to test algorithms' performance when applied to video captured in real-world settings. They found that when asked to analyze footage of individuals walking through a sporting venue, median false-negative rates for the leading algorithm varied from a **low of 13 percent all the way up to 64 percent**, depending on camera placement.[4]

This is far too high of an error rate for this type of deployment to be reliable in practice, but similar tests in more constrained settings have provided better results. For instance, in 2020, the U.S. Department of Homeland Security's Maryland Test Facility (MdTF) tested 60 combinations of commercial face acquisition and matching systems as part of their **2020 Biometric Rally**. The test involved sending volunteers through an environment similar to what may be found at a border checkpoint or stadium ticket station. Participants approached one at a time in an orderly manner into an area where a camera was positioned to capture a clear head-on photo. The leading algorithm achieved a 99.7 percent true positive rate during the simulation, while the median algorithm achieved a true positive rate of 93 percent. When taken together with the results from NIST's FIVE, this shows that while many real-world settings will likely continue to pose significant challenges for facial recognition systems, a combination of technical improvements and careful control of image capture environments may soon make it possible for these systems to be deployed in certain real-world settings with a high degree of reliability.

Other than image quality, the other prime factor impacting facial recognition performance is the algorithm quality. In their most recent report, NIST **emphasized** that "recognition accuracy is very strongly dependent on the algorithm and, more generally, on the developer of the algorithm. False negative error rates in a particular scenario range from a few tenths of one percent to beyond fifty percent," depending on the vendor being evaluated. In NIST's FIVE, for example, the leading algorithm had an observed error rate of 13 percent on the best-quality video stream, but the error rate for the median algorithm on that same footage **jumped to 60 percent**.[5] For the lowest-quality stream, the median algorithm had a false negative rate of **92 percent**, meaning that over half of the algorithms missed more than 9 out of 10 matches.[6] Even in MdTF's more recent evaluation, researchers found that only 4 of the 60 combinations of acquisition and matching systems they tested achieved the target goal of a 99 percent true positive rate.

[3] COGENT-004 (0,2] and DEEPSEA-001 (14,18] FNIR at identification for FPIR = 0.001. Table 5.

[4] M32V FNIR on identification at low 6ft placement and door 8ft placement (near field) on dataset P.

[5] N31V FNIR on identification at high 6ft placement (near field) on dataset P.

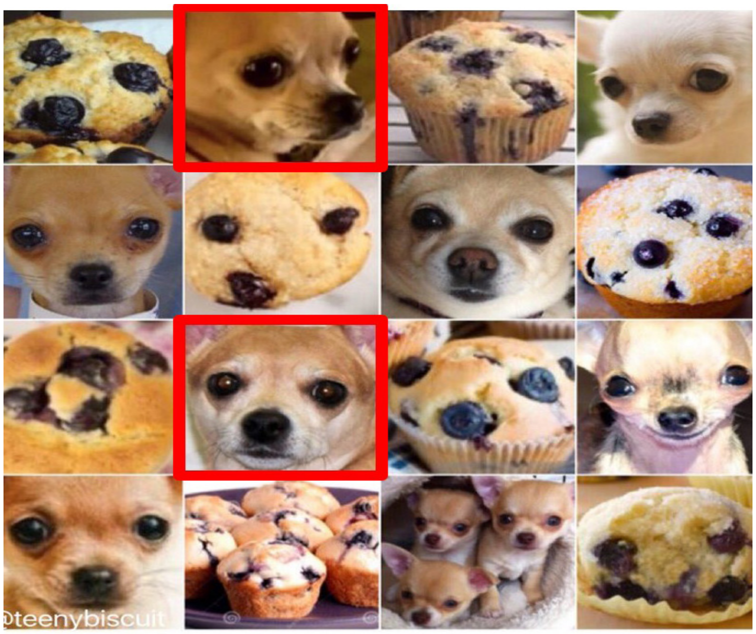
[6] J31V FNIR on identification at door 8ft placement (near field) on dataset P.

This shows that achieving reliable performance requires *both* high-quality images and high-quality algorithms. While a few leading vendors have managed to develop facial recognition algorithms that can achieve very high accuracy in controlled conditions, the average facial recognition provider still struggles to achieve similar reliability, and even the best still experience issues when deployed in unconstrained settings. This makes it difficult to come to broad conclusions, either positive or negative, about the accuracy of facial recognition systems, as the operational performance is so heavily dependent on the vendor and the context of deployment.

What can be said is that a facial recognition system is much more likely to have high performance in situations where subjects are cooperative and aware their image is being captured, where camera placement allows for good lighting and positioning, where the photos being matched to are of high quality, and where the algorithm is sourced from a leading vendor. This shows that policymakers and operators need to consider the circumstances of deployment and the algorithm being used to fully understand the potential risks of a deployment.

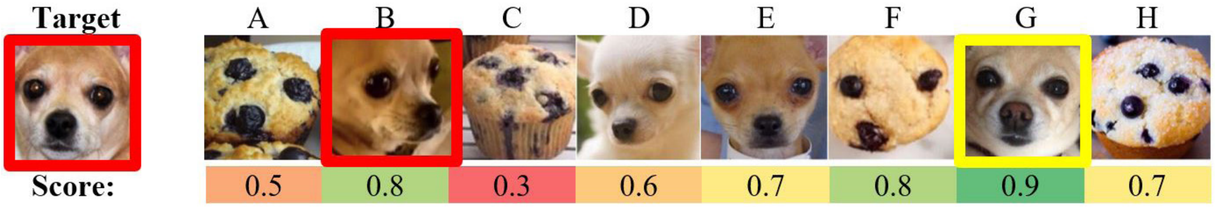
Understanding Comparison Thresholds

The previous section established that there are two kinds of errors a facial recognition system can make: false negatives and false positives. Understanding the relationship between these two kinds of errors is important for the operators responsible for adjudicating matches and for policymakers considering the potential impacts of facial recognition deployments. In general, it is difficult to reduce the rate of either false positives or false negatives without raising the other. To understand why, consider the following set of pictures featuring Chihuahuas and blueberry muffins composed by **Twitter user @teenybiscuit**:



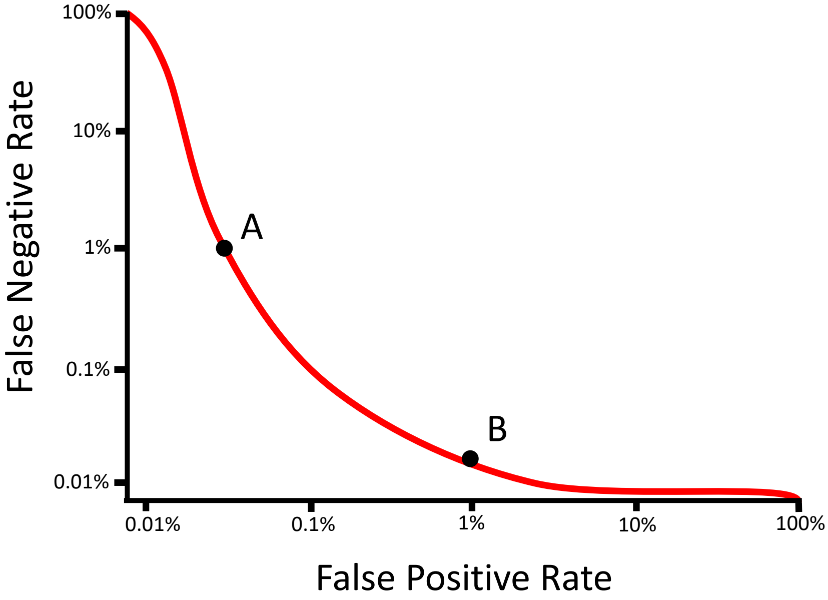
Source: Karen Zak, Twitter Post, March 9, 2016, 4:40 p.m., https://twitter.com/teenybiscuit/status/707727863571582978.

Imagine that someone wanted to use this series of images to test out a facial recognition system for dogs. The test was whether the system could correctly determine that the two images highlighted in red are of the same dog. One of the highlighted images was selected and input into the system as the target for matching to run the test. After processing each of the other images, the system output the following similarity scores (ranging between 0 and 1), indicating how similar it calculates each image is to the target:



We know that image B is actually the correct match, but we can see from these results that the system believes that the most similar image is actually G. If the system had been set up to only report back images with scores of at least 0.9—minimizing the risk that it would return false positive matches—then it would have rejected the correct match—a false negative. If, however, the system was instead set to report back every match with scores of at least 0.8, its results would include the correct match (a true positive) but also two false positives. And one of those would have been a muffin.[7]

From this, we can see that imposing high similarity score thresholds can sometimes mean that the correct match ends up being rejected. But by lowering the threshold, we let in more false positives. For any given algorithm, the trade-off between false negative rates and false positive rates as this threshold is changed can be depicted using a graph called a detection error tradeoff (DET) curve. The figure below depicts an example of what the DET curve for a hypothetical algorithm may look like. It is important to note that the exact shape of this curve will vary for each algorithm.



This curve represents the different combinations of false positive and false negative error rates that an algorithm can achieve. Depending on the similarity threshold used, an algorithm can theoretically sit at any point along this curve. Instituting a high similarity threshold (like 0.9 in the chihuahua example above) may lead to the algorithm operating at point A on the graph above. At this point, fewer than 1 in 1000

[7] This example is for illustrative purposes only and should not be taken as representing the actual performance of modern computer vision systems.

comparisons between images of two different people will yield a false match. However, 1 in 100 comparisons between two images of the same person will be incorrectly rejected.

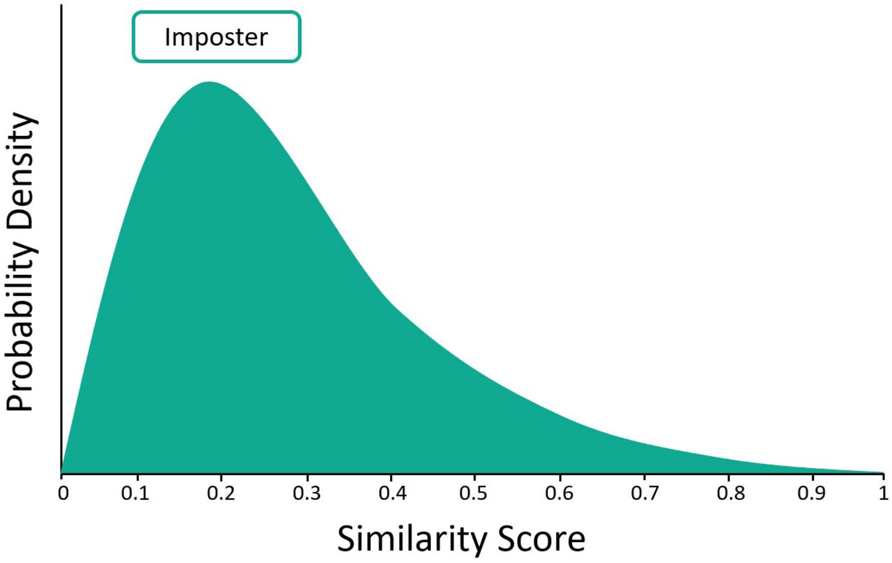
Alternatively, in deployments where the risks of false negatives are greater than the risk of false positives, operators may decide to institute a lower threshold, leading to the algorithm operating at point B. Here, the relationship would be reversed, with 1 in 100 comparisons between images of two different people leading a false match, and fewer than 1 in 1000 comparisons between two images of the same person being rejected.

There is no single correct answer about what threshold a facial recognition system should be set to. It depends entirely on what the technology is being used for. In some cases, like terrorist screenings at military bases, the risks resulting from a false negative (failure to recognize a terrorist suspect) are often much higher than the harms resulting from a false positive (pulling additional innocent individuals out for secondary screening). Here, operators will likely prefer lower similarity thresholds. Low thresholds may also be appropriate in some commercial applications, like when using facial recognition in retail stores to sign customers into customer loyalty programs, as the likelihood and risks of someone logging into another person's account are negligible.

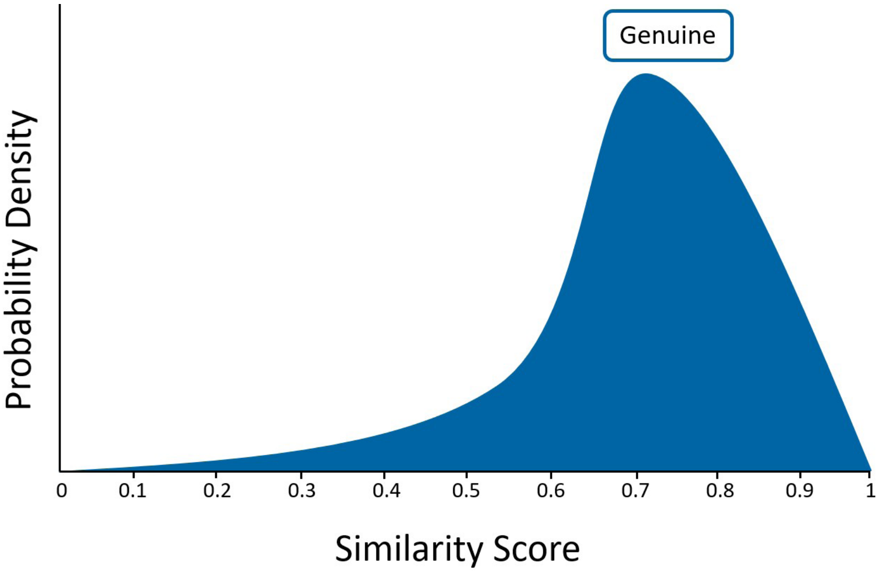
In other circumstances, false positives may pose a greater risk than false negatives. The most obvious example of this kind of situation is when facial recognition is used by law enforcement to generate investigative leads. Here, acting on erroneous matches could lead to the arrest of innocent people. However, it is important to note that in most law enforcement deployments in the United States, facial recognition is used to produce a list of potential leads for investigators to review rather than as a way of singling out a particular suspect for arrest. When used this way, there is an opportunity for operators to institute additional layers of safeguards to identify false positives before those errors lead to negative consequences.

Understanding Similarity Scores

Similarity scores form the basis for setting comparison thresholds, but they are also important to understand due to the way they are often misinterpreted by operators. Recall the previous hypothetical example of chihuahuas and blueberry muffins, where the system was set to return a score between 0 and 1 describing the degree of similarity between two images. These similarity scores are often misunderstood as translating directly to percentages (i.e., a similarity score of 0.9 means there is a 90 percent chance that the match is correct). But this is not what the score actually represents. The figures below represent a simplified graph of all the similarity scores calculated by a hypothetical system as it tests different pairs of faces:

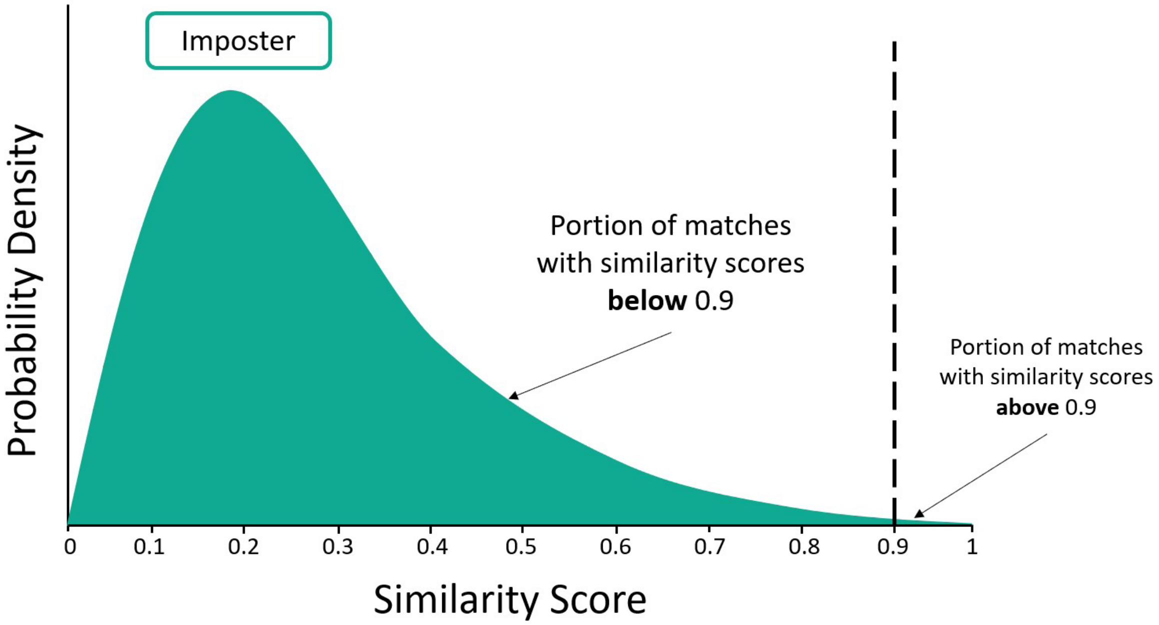


In this graph, the green curve represents the he distribution of similarity scores generated when a target image is compared agaisnt the face of a different person (referred to as an imposter). As you can see, the large majority of these cases yielded low similarity scores, but a small number of test cases yielded higher scores due to the fact that some imposters may closely resemble the target.



In contrast to the first graph, the blue curve in the second graph shows that a majority of comparisons between genuine matches yield high similarity scores, though a few produce low scores due to issues with quality, aging, or other factors. Importantly, each developer has a different distribution associated with their algorithm, so even though they may present their similarity scores in a similar form, a given score means something different for each vendor.

These probability distributions allow us to make statements about the chances of particular events occurring. For example, we can find the probability that a random imposter will return a similarity score of 0.9 or above by dividing the area of the green curve to the right of 0.9 by the area of the green curve to the left of 0.9.



As we can see, this probability will be very small—perhaps 1 percent. However, this is different than saying that matches yielding a similarity score of 0.9 have a 99 percent chance of being correct. To understand why, we must explore a common statistical mistake known as the "prosecutor's fallacy."

As we all know, all dogs have four legs (with a small number of exceptions due to accidents and genetics). From this, we can make a straightforward statistical statement about the probability of a dog having four legs in the form of "given A, what is the probability of B?" We can represent "being a dog" as condition A, and "having four legs" as condition B:

> If A is true, what is the probability of B? → "If an animal is a dog, then the probability it has four legs is 99%."

Simple enough. But then we change it to take the converse form:

> If B is true, what is the probability of A? → "If an animal has four legs, then the probability it is a dog is 99%."

This is obviously wrong, demonstrating that you cannot simply reverse the order of such a statement and be sure it will remain true. However, there are other times when this mistake may not be as obvious. For example, one could imagine running facial recognition on two images and getting back a similarity score of 0.9. Now say that according to that vendor's particular distribution, there is only a 1 percent chance of a similarity score of 0.9 occurring if the target image and test image are from different people. From here, we can redefine A to represent "the facial recognition match is incorrect" and B to represent "the system returns a similarity score of 0.9":

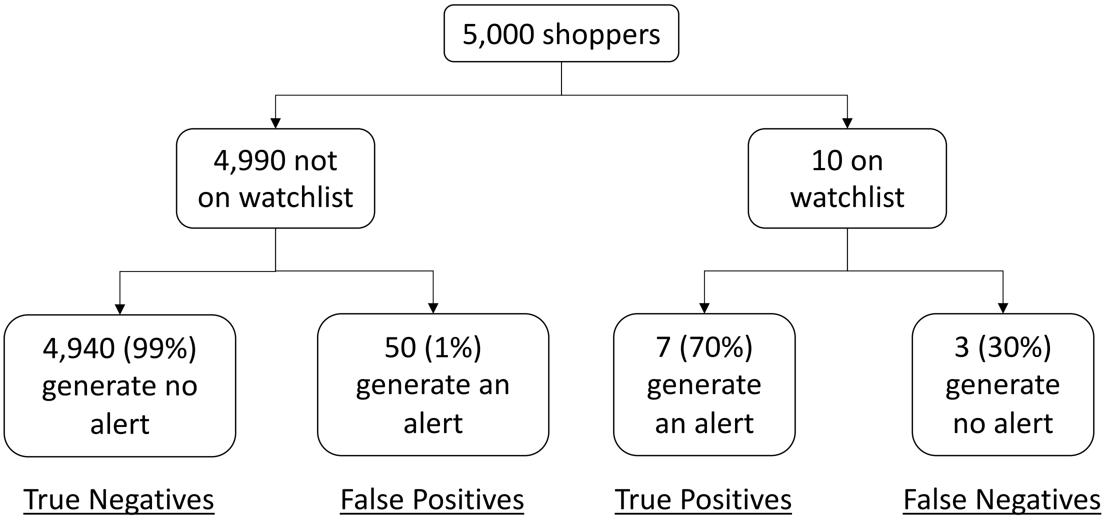
> If A is true, what is the probability of B → "If the facial recognition match is incorrect, then the probability that the system returns a similarity score of 0.9 is 1%."

This is the precise statistical meaning of a facial recognition similarity score, as we explained above. But if we then change it to the converse form like the former example, it becomes:

> If B is true, what is the probability of A → "If the system returns a similarity score of 0.9, then the probability the facial recognition match is incorrect is 1%."

In this example, the converse is not obviously incorrect as in the first case, but it is still just as wrong. This is known as the prosecutor's fallacy because of the way this flaw in statistical reasoning is often used to misrepresent the meaning of forensic evidence. In reality, the likelihood that a given match with a similarity score of 0.9 is incorrect could be much higher than 1 percent. To illustrate this, imagine a trial run of a face-based shoplifter detection system that worked by comparing the faces of each person who entered a mall against a watchlist of people who had been caught shoplifting there over the past two years.

Imagine that out of 5,000 people who visit on a particular day, 10 of them are former shoplifters who are included in the mall's watchlist. Now, let us say that the facial recognition system is accurate enough to have a 70 percent chance of identifying an imposter when they walk in (a 30 percent false negative rate). Let us also say that given the comparison threshold used by the operator, the system has a 1 percent chance of mistakenly matching an innocent individual to a person on the watchlist (a 1 percent false positive rate). From this, we can predict the following will happen:



From this, we can see that the system would likely generate around 57 alerts in an average day, of which only 7 (12 percent) would be correct. This is not necessarily a problem if the operators are made aware of these statistics and the fact that the large majority of alerts the system generates are likely to be false. In this case, human reviewers can be careful to double-check alerts to make sure they only act on clear matches. The problem emerges when the vendor of such a system reports that their algorithm is set only to report matches it is "99 percent confident in," and then operators who do not know the difference assume every alert has a 99 percent chance of being a genuine match.

This is why operator training and match presentation is so important to managing the risks of facial recognition systems. Operators must be aware of how their systems actually operate if they are to make informed judgments about how to adjudicate potential matches. Developers should also take this into consideration and not build user interfaces that present operators with misleading information about the meaning of the scores they return.

Photo by Kevin Frayer/Getty Images

This is not how facial recognition actually works.

Is Facial Recognition Biased?

One of the most common criticisms of facial recognition is that it has different accuracy rates for different demographic groups, leading to risks that members of certain groups are more likely to suffer the consequences of being misidentified, particularly in the law enforcement context. The most thorough investigation of this disparity was completed by NIST in 2019. Through their **testing**, NIST confirmed that a majority of algorithms exhibit demographic differences in both false negative and false positive rates.

Using curated images, NIST found that, in general, Asians, African Americans, and American Indians had higher false positive error rates than white individuals, women had higher false positive rates than men, and children and the elderly had higher false positive rates than middle-aged adults.[8] Differences in false positive rates are generally of greater concern to privacy advocates, as there is usually greater risk in to subjects in being misidentified than in being incorrectly rejected by a facial recognition system. However, in some applications such as access control, there may actually be a greater risk in missing the correct identification of bad actors, leading to greater interest in the false negative rates. NIST found that demographic factors had a much larger effect on false positive rates (where differences in the error rate between demographic groups could vary by a factor of 10 or even 100) than false negative rates (where differences were generally within a factor of 3 depending on the threshold value for a positive match).

However, NIST also came to several encouraging conclusions. The first is that differences between demographic groups were far lower in algorithms that were more accurate overall. This means that as facial recognition systems continue to improve, the effects of bias will be reduced. Even more promising was that some algorithms demonstrated no discernible bias whatsoever, indicating that it may be possible to eliminate bias with the right algorithms and development processes.. One of the most important factors in reducing bias appears to be the selection of training data used to build algorithmic models. If algorithms are trained on data sets that contain very few examples of a particular demographic group, the resulting model will be worse at accurately recognizing members of that group in real-world deployments. NIST's researchers theorized that this may be the reason many algorithms developed in the United States performed worse on Asian faces than algorithms developed in China. Chinese teams likely used training data sets with greater representation of Asian faces, improving their performance on that group.

In addition to training data, the quality of the images being captured also has a large influence on demographic differences. An **assessment** of 11 commercial facial recognition systems by MdTF found that dark skin was associated with lower similarity scores but that the skin reflectance of the subject was a better predictor of accuracy differences than the self-reported race of the subjects. This may indicate that higher-quality cameras and better image capture setups could make a large difference in eliminating demographic bias by improving operators' ability to take clear images of dark-skinned individuals. Similar to NIST, the MdTF found that the most accurate algorithm overall had an almost negligible demographic effect, further supporting the conclusion that improvements in algorithm quality will gradually reduce bias in these systems.

It is important to note that popular reporting about the issue of facial recognition bias often conflates facial recognition with face characterization. In particular, there has been significant media attention paid to the 2018 **Gender Shades** study, which tested three commercial gender classification systems and found that darker-skinned females were misclassified as male up to 34.7 percent of the time, compared to light-skinned men, who were only misclassified as women up to 0.8 percent of the time. This obviously represents a significant discrepancy, and these findings have helped drive focus and attention on the issue of algorithmic bias in recent years.

[8] The racial categories used by NIST were based on the FBI's "Electronic Biometric Transmission Specification Technical and Operational Update (TOU)" 10.0.9, May 22, 2018

However, it is important to remember that this finding relates to face characterization systems, not face recognition systems. Popular reporting frequently elides this distinction and implies that this level of bias also applies to facial recognition systems. While many of the underlying causes of bias in facial recognition and facial characterization are similar (e.g., lack of representative training data, poor lighting, and image capture), finding a particular magnitude of bias in one does not prove that the same magnitude of difference is present in the other. Particularly given that face characterization is a much younger and less well-developed field of study than facial recognition, it is to be expected that the technology will have worse performance. The Gender Shades study is an excellent resource for policymakers thinking about the risks associated with deploying gender characterization technology. But when discussing face recognition, policymakers should base their deliberations on the risks posed by demographic differences in accuracy rates, primarily on evaluations such as those conducted by NIST and MdTF, which specifically tackle recognition technologies.

Conclusion

As policymakers and operators work to develop new rules and regulations to govern the use of FRTs, it is important that these efforts be grounded in facts about how the technology works. Facial recognition is usually discussed only in the context of its most dystopic applications, but in reality, it is a multifaceted tool that can be applied to a range of different problems, from signing a user into their phone to generating leads in law enforcement investigations. In the majority of these cases, facial recognition is used to aid human decision-making rather than replace it.

This human oversight helps to mitigate the risk of errors which are still a major problem for many real-world deployments. While technical improvements will help to improve facial recognition performance and reduce demographic differences in accuracy rates, facial recognition will never be perfect. To manage risks, operators must understand how system performance can be affected by the conditions of a deployment, select appropriate confidence thresholds to manage the trade-off between accuracy and risk, and adopt safeguards that account for how theoretical performance translates to real-world impacts.

Operators, policymakers, and the public must also understand the difference between facial recognition and facial characterization. This crucial distinction is too often blurred, leading to criticism that is misleading or incorrect. An improved understanding of the issues covered in this report will help ensure this technology can be deployed safely in a way that allows us to capture its benefits while managing the risks. ■

William Crumpler *is a former research associate with the Strategic Technologies Program at the Center for Strategic and International Studies (CSIS) in Washington, D.C.* ***James Andrew Lewis*** *is senior vice president and director of the Strategic Technologies Program at CSIS.*

This report was funded in part by the Department of Homeland Security as part of its homeland security mission to defend the homeland while upholding our nation's values.

This report does not constitute the official position of the Department of Homeland Security. Any questions can be directed to the Office of Public Affairs at the Office of Biometric Identity Management.

This report is produced by the Center for Strategic and International Studies (CSIS), a private, tax- exempt institution focusing on international public policy issues. Its research is nonpartisan and nonproprietary. CSIS does not take specific policy positions. Accordingly, all views, positions, and conclusions expressed in this publication should be understood to be solely those of the author(s).