ID.me

# Promoting Access, Equity, and Inclusion

## With AI and Digital Identity

# Introduction

As services and transactions increasingly move online and the economy becomes ever more digital, we have an enduring obligation to ensure everyone has the opportunity to participate. ID.me makes that real with our "No Identity Left Behind" initiative, a fundamental commitment to equity and access. Equity is why we do what we do. This paper shares insights on the application of AI and facial recognition in identity proofing at NIST SP 800-63-3 IAL2. IAL2 is the federal standard that defines required and suggested controls for authenticating consumers for high-risk services. Examples include:

- **Authenticating individuals for access to taxpayer services**
- **Enabling individuals to apply for unemployment benefits**
- **Empowering consumers to access health care records in the public and private sectors**

Equity with respect to AI and facial recognition can be difficult to parse because there are many different applications of the technology and scrutiny on how law enforcement, in particular, uses it. Still, it is crucial to unpack how AI affects individuals from various communities and demographics as they attempt to access vital government programs. Given that AI and facial recognition can automate many workflows – enabling faster service delivery – the empirical data and truth must win over false perception.

Findings from a 2019 NIST report on facial recognition are important for policymakers because those findings relate to NIST SP 800-63-3 IAL2. Understanding how the leading facial recognition algorithms affect equity and access in the context of NIST 800-63-3 can help policymakers understand if IAL2 requirements are equitable. That is vital to ensure optimal policies for equity and access and so the public understands the controls that are used.

### The best available research and data on those topics paint a clear and hopeful picture:

- **The leading algorithms show extremely high accuracy across all demographics in IAL2 flows**
- **ID.me internal tests across 15,468 images show no detectable bias tied to skin type**
- **Mitigating controls – such as human reviewers and in-person verification – control for any potential bias**

The ID.me NIST IAL2 solution uses leading algorithms as validated by NIST and NIST-accredited laboratory testing. ID.me also employs two sets of human reviewers to check the technology's decision when AI denies access in the self-serve flow. That hybrid approach enables the leading algorithms, which are more accurate and less biased than trained humans,[1] to streamline access while mitigating any risk of bias.

## The remainder of this paper goes beyond the headlines, unpacks the science, and explains:

- ✔ **An overview of the 2019 NIST Face Recognition Vendor Test (FRVT)**
- ✔ **The difference between 1:1 face match and 1:many facial recognition and why it matters**
- ✔ **The critical difference between a false positive and a false negative**
- ✔ **Confusion about facial recognition and pass rates by gender**
- ✔ **NIST's findings on the highest-quality algorithms**
- ✔ **How ID.me uses those findings and Trusted Referees to enhance equity in identity verification**

1 Crumpler, William, How accurate are Facial Recognition Systems - and Why Does It Matter?, Center for Strategic & International Studies https://www.csis.org/blogs/technology-policy-blog/how-accurate-are-facial-recognition-systems-%E2%80%93-and-why-does-it-matter

# An Overview of the 2019 NIST FRVT

In 2019, NIST conducted a study of more than 189 commercial algorithms from 99 developers to quantify the accuracy of facial-recognition algorithms for different demographic groups. Notably, many of those algorithms were immature and submitted by universities for research purposes. Test results from those algorithms should not be conflated with the performance of leading algorithms or algorithms actually used by IAL2 vendors.[2]

The results were based on a dataset of more than 18 million images of 8.5 million individuals. Key findings include[3]:

- **Algorithms perform differently:** The results show a wide range in accuracy across developers. The best performers produce "many fewer errors" than less-mature algorithms. Mature algorithms can therefore be expected to have smaller demographic differentials.
- **Demographic effect is vanishingly small:** False negatives – when a legitimate person's selfie fails to match a reference photo of his or her face – occur at extremely low rates across demographic groups. That is particularly important because a false-negative error would deny a legitimate person access.
- **Leading algorithms perform more equitably:** The best facial-recognition algorithms perform more equitably across demographic groups for 1:1 face match in scenarios when a valid user is attempting to pass. False-negative errors, which block valid people, are usually remedied on a second attempt.
- **Confusion about bias abounds:** Media reports and even university studies often fail to use precise terminology and, as a result, negatively skew the public discourse. For example, studies on gender, which related to face-classification algorithms, were falsely conflated with facial recognition, which looks for similarity.

What does that mean to organizations seeking to leverage those technologies to provide secure, equitable services? In short, they should understand how performance varies across types of algorithms (for example, 1:1 vs. 1:many), they should adopt only the highest-performing algorithms, and they should take action to mitigate known and potential performance limitations and errors. The sections that follow provide insights into how to take those actions to increase access, equity, and inclusion in digital identity.

2 IAL2 vendors should disclose the specific algorithms they are using to certifying bodies so they can be evaluated for equity and inclusion.

3 Grother, Patrick; Ngan, Mei; and Hanaoka, Kayee. Face Recognition Vendor Test (FRVT), Part 3: Demographic Effects, 2019, National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf

# The Difference Between 1:1 Face Match and 1:Many Facial Recognition and Why It Matters

Let's clarify why there are more errors tied to more complex use cases. 1:1 face match is equivalent to an airport agent comparing your face to the photo on your government ID card. 1:many facial recognition is equivalent to giving your picture to the same agent, putting him on stage at a rock concert, and asking him to pick your face out of the crowd. With millions of possible matches, the challenge of finding the right face increases measurably.[4] Face match tied to NIST IAL2 deals specifically with 1:1 matching. The goal is to avoid a false negative so legitimate people are able to gain access. An additional goal is to avoid a false positive so an identity thief is unable to claim a different person's identity. That is a simplistic use case in the context of advanced technology.

1:Many Facial Recognition

1:1 Face Match

With more than 129 million Android smartphones and 113 million iPhones in use in the U.S.,1:1 face match is already widely adopted by tens of millions of Americans. It has been proven at scale. *False negatives* occur when the technology fails to match the same person from the FaceID enrollment photo to the image captured during a specific attempt to unlock the phone. Apple and Android manufacturers allow for additional attempts and then prompt the user for a PIN if repeated attempts fail. While that content isn't covered by the 2019 NIST report directly, it provides a helpful frame to interpret the results of the study and how leading companies introduce additional controls to provide access pathways in the event AI doesn't perform as intended.

4 NIST benchmarks FPIR for 1M+ databases at 0.001, which is much higher than 1e-6 for 1:1,. but still excellent for many use cases.

# The Difference Between False Positives and False Negatives

False-positive errors occur when two faces look similar but do not belong to the same person. Those errors are often embarrassing when humans make them in social interactions, such as mistaking a stranger for a friend. A false-negative scenario might involve failing to recognize an old friend you went to school with years ago.

False positives are much more common in 1:many facial-recognition scenarios. They are far less common in 1:1 face matching. After all, what are the odds that a person who steals your wallet looks just like you?

**False Negative – Algorithm Fails to Match Two Images of the Same Person's Face**

**False Positive – Algorithm Incorrectly Matches Two Faces That Aren't the Same Person**

When verifying identity for government benefits, false negatives would be associated with denying access to a person who is the same as in the government ID photo. The NIST report shows that false-negative errors are vanishingly small across demographic groups. To the extent false-negative errors occur across all algorithms, false-negative errors are actually lower in darker skin tones for 1:1 matching under certain conditions. Keep in mind, false negatives can often be remedied by trying a second time, as NIST notes.

The errors related to false positives are most relevant for fraud and unauthorized access. Those errors would not relate to legitimate people getting blocked from their rightful benefits, but rather to a criminal gaining unauthorized access. The NIST report notes "false positive differentials are much larger than those related to false negatives and exist broadly, across many, but not all, algorithms tested." While that is relevant for 1:many anti-fraud scenarios, **the false-negative rate is the key metric in 1:1 identity verification as it deals with blocking a valid person.**

**The NIST report highlights three additional findings related to error rates:**

- **False negatives** are often remedied by the user attempting a second time
- **False-negative rates** are extremely low across demographic groups
- **False-negative errors** tend to be algorithm specific

## The Difference Between Face Classification and Facial Recognition and How They Perform Across Genders

The 2019 NIST report addressed confusion in the market about facial-recognition versus facial-classification algorithms as they relate to pass rates across genders. The excerpt from the NIST report that highlights the confusion, and how it affects perceptions of bias, follows with bolding added for emphasis by ID.me:

"Over the last two years there has been expanded coverage of face recognition in the popular press. In some part this is due to the expanded capability of the algorithms, a larger number of applications, lowered barriers to algorithm development, and, not least, reports that the technology is somehow biased. This latter aspect is based on Georgetown and two reports by MIT. The Georgetown work noted prior studies articulated sources of bias, and described the potential impacts particularly in a policing context, and discussed policy and regulatory implications. **The MIT work did not study face recognition, instead it looked at how well publicly accessible cloud-based estimation algorithms can determine gender from a single image. The studies have been widely cited as evidence that face recognition is biased.**

This stems from a confusion in terminology: Face classification algorithms, of the kind MIT reported on, accept one face image sample and produce an estimate of age, or sex, or some other property of the subject. **Face recognition algorithms, on the other hand, operate as differential operators:** They compare identity information in features vectors extracted from two face image samples and produce a measure of similarity between the two, which can be used to answer the question 'same person or not?'. Face algorithms, both one-to-one identity verification and one-to-many search algorithms, are built on this differential comparison."[5]

5 Grother, Patrick; Ngan, Mei; and Hanaoka, Kayee. Introduction, page 14. Face Recognition Vendor Test (FRVT), Part 3: Demographic Effects, 2019, National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf

The NIST report goes on to compare false-negative rates of the 52 most accurate recognition and matching algorithms to the classification algorithms in the MIT study: The best algorithms "almost always gives (sic) false-non-match rate (FNMR) below 1%. These error rates are far better than the gender-classification error rates that spawned widespread coverage of bias in face recognition. In that study, two algorithms assigned the wrong gender to black females almost 35% of the time. The recognition error rates here, even from middling algorithms, are an order of magnitude lower. Thus, to the extent there are demographic differentials, they are much smaller than those that (correctly) motivated criticisms of the 2017-era gender classification algorithms."[6]

6 Idib. Page 54.

# NIST Findings on the Highest-Quality Algorithms

NIST found leading algorithms to be exceptionally accurate, with far fewer errors and smaller differentials across demographic groups. Breaking that down a bit further, leading algorithms have false-negative rates that are "usually low with average demographic differentials being, necessarily, smaller still."[7] For false positives, results were a bit more nuanced. NIST found that "false positive differentials (across demographics) are much larger than those related to false negatives across many, but not all, algorithms tested."[8] That means only a few algorithms were found to have low error rates and low demographic differentials across both false negatives and false positives. For that reason, it is critical that conversations around accuracy account for the type and quality of the algorithm, as determined by NIST FRVT rankings.

After a rigorous screening and selection process, ID.me partnered with Paravision, which has been repeatedly benchmarked by NIST as a global accuracy leader and in internal testing has demonstrated 99.8% transactional pass rates (2 in 1,000 false-negative rates) and 1 in 100,000 false-positive rates across a wide range of demographic groups. Generally, NIST found that false-negative errors are typically remedied by repeating an image-pair comparison, for example uploading a new selfie. That means if an individual fails to match on the selfie step on the first attempt, a second attempt typically passes. Those findings are consistent with how ID.me uses Paravision and actual pass rates in an applied setting.

7 Ibid. Page 10.

8 Ibid. Page 5.

# How ID.me Uses the Findings and Trusted Referees to Enhance Equity in Identity Verification

ID.me is committed to our "No Identity Left Behind" initiative. We built our identity-verification products by combining best-in-class technology with human-in-the-loop relief valves. From a technology standpoint, ID.me uses the best face-matching and presentation-attack detection (PAD), or face liveness, capabilities available in the market. We monitor associated error rates to detect any potential bias and to improve pass rates.

The ID.me face-match step has a 98.9% pass rate per user and a 98.5% pass rate per transaction in our self-serve flow, which includes variables such as image quality, lighting, and skew. Improvements to user-experience copy and to FAQ pages have increased the rate from 95% in March 2021. That gain shows the power of usability research, language accessibility, and data feedback loops to improve overall accessibility over time. Keep in mind, that pass rate is likely artificially low because it does not account for fraudulent attempts.

ID.me always implements secondary and tertiary controls to ensure no user is blocked by a false negative. To do so, we direct any user who fails the 1:1 face-match step to join a video chat with a human agent, a Trusted Referee. Trusted Referees primarily support users who encounter challenges, such as thin credit files and recent name changes, to attain verification. They are also available for face-match backup. In addition, we layer in a team to review automated decisions in real time to ensure two levels of human review in the online flow.

ID.me has verified more than 2.8 million people through Trusted Referees. Recently, ID.me also partnered with Sterling to make in-person identity verification available at 650 retail locations across the country. New Jersey was the first state to adopt that identity verification method.[9] In so doing, ID.me offers multiple relief valves or escape hatches to ensure there is always a path forward for everyone. We are committed to a policy of "No identity left behind."

9 GCN Staff. NJ offers in-person ID verification for online services: https://gcn.com/cybersecurity/2021/11/nj-offers-in-person-id-verification-for-online-services/316338/

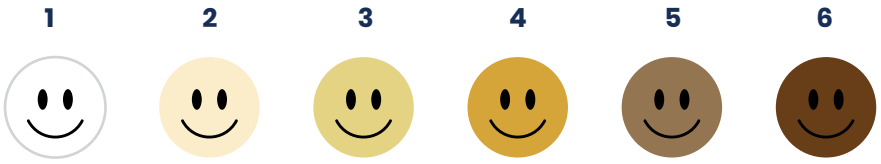
# Technology Performance

ID.me uses Paravision, the top-ranked 1:1 face-matching algorithm developed in the U.S. and a leading algorithm globally across all leading benchmarks:

- **1:1 Verification**
- **1:Many Verification**
- **Face-Mask Effects**
- **Paperless Travel**
- **Quality Assessment**

When it comes to false negatives, the algorithms in use by ID.me do not exhibit operationally significant differentials across demographics. It is also important to note that technology performance improves with each year. The leading algorithms, which were already equitable for 1:1 access in 2019, have advanced further over time and have been tested for performance gains against multiple demographic groups.

In March 2021, ID.me performed tests to look for bias related to face match and skin tone. We picked the Social Security Administration for analysis as a broadly representative government agency that is not a target for fraud like state unemployment agencies. We then pulled a randomized sample of 627 individuals who had failed the face-match step. We used the Fitzpatrick skin type framework to classify individuals: 1 being the lightest and 6 being the darkest.




**Fitzpatrick Skin Type Scale**

A regression analysis did not yield any P values lower than .05 to correlate a given skin tone to face-match failure. We also ran a Chi-Square test for categorical variables and proportion tests for significant differences in proportions for group and reason while controlling for sample size. Neither test presented evidence of a relationship between skin type and failure reason. **See Appendix A for the March 2021 test results**.

In December 2021, ID.me performed additional tests to look for bias related to face match and skin tone per the Fitzpatrick Skin Type Scale. We picked the IRS as a separate agency that is also broadly representative and not an extreme target for fraud outside of tax season. We used 15,468 labeled images, collected in two sample sets for separate tests. The first tests were run to evaluate the selfie 1:1 match for any correlation between the Fitzpatrick Scale number and the rates of selfie-match 1:1 failures. The second set was used to run the same tests on liveness data. Both samples passed

the Chi Square Independence Test, indicating selfie-match and liveness-failure rates were independent of skin type value on the Fitzpatrick Scale. **See Appendix B for the December 2021 test results**.

Paravision performed a 1:1 face-match demographic assessment using a highly diverse set of 70,000 face images against known match images and nonmatch images. The two metrics Paravision measured the dataset against are false-nonmatch rate (rate at which it should have matched two images and didn't) and false-match rate (rate at which it matched two images when it shouldn't have). Within every ethnicity represented in the dataset, Paravision achieved a false-nonmatch rate of less than 2 in 1,000 and a false-match rate of less than 1 in 100,000. That means that out of the dataset and across all skin types and genders, 2 out of 1,000 should have been matched and weren't and 1 out of 100,000 were matched and shouldn't have been.

For PAD, to ensure the selfie submitted is actually that user's face, ID.me partners with iProov, a leading vendor that has been independently tested by an evaluating body accredited by the NIST National Voluntary Laboratory Accreditation Program. The laboratory's accreditation also includes ISO 30107-3, the international standard that governs PAD. The UK National Physical Laboratory audited iProov's performance data, found the solution conformant with ISO 30107-3, and concluded that performance is "state of the art." In a trial conducted in 2021 under controlled conditions for a UK government agency with a diverse set of 500 individuals, balanced for ethnicity, age, and gender, 99.91% of users successfully completed the face-liveness process and passed.

iProov's service is already deployed in different regions across the world. As a result, performance variation can be compared across countries and regions with varying demographics. Signals of bias, including age, gender and ethnicity, are periodically calculated to seek out any potential inequities and enable remediation. In South Africa, like-for-like performance rates are not different from other regions, with no detectable differential pass rate that might negatively affect different demographic or ethnic groups. In Singapore, iProov passed bias testing administered by Govtech. **See Appendix C for more detailed information about iProov's PAD performance during testing through a NIST-accredited lab**.

# The Role of Humans in the Loop

ID.me uses AI and facial recognition in an ethical manner. We believe every system must be built in a resilient manner to detect potential biases that might have otherwise gone unobserved and to ensure there is a real-time to near real-time path forward for any affected user, regardless of the reason. As a result, we employ teams of trained human agents as relief valves – just like a smartphone will prompt the user for a PIN to unlock the device if too many selfie attempts fail. There is always a way forward and feedback loops too.

Peer-reviewed scientific studies[10] show that leading algorithms are more accurate than even forensic examiners, who specialize in facial comparison, at comparing two human faces. Additionally, human reviewers are subject to bias. That bias is inherently harder to standardize and control because it varies by each individual.

Studies emphasize that the best face-match results are achieved by fusing computer-based and human-driven facial recognition. That is exactly what ID.me does. If a computer algorithm cannot conclude that a given person matches a given photo in a government-issued ID, ID.me will then invoke human-based recognition by inviting the ID.me member into a video chat with one of more than 1,000 U.S.-based, specially trained Trusted Referees.

According to NIST 800-63-3, Trusted Referees "assist in the identity proofing and enrollment for populations that are unable to meet IAL2 and IAL3 identity proofing requirements or otherwise would be challenged to perform identity proofing and enrollment process requirements." Examples of populations include:

- **Disabled individuals**
- **Elderly individuals**
- **Homeless individuals**
- **Individuals with little or no access to online services or computing devices**
- **Unbanked individuals and those with little or no credit history**
- **Victims of identity theft**
- **Children under 18**
- **Immigrants**

10 "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," https://www.pnas.org/content/pnas/115/24/6171.full.pdf

Our team of Trusted Referees represent all ethnicity groups in the U.S. census. We meet or exceed the level of minority representation in the U.S. population for the four largest nonwhite ethnicity groups: Black or African American, Hispanic or Latino, Asian, and two or more races. To further enhance equity, our Trusted Referees can verify individuals in 16 languages: English, Spanish, Haitian Creole, Korean, Arabic, Mandarin, Cantonese, Hindi, Farsi, Wolof, Nepali, Mandingo, Punjabi, Urdu, Russian, and American Sign Language.

We believe that combining a talented pool of service-minded Americans with best-in-breed technology is the best way to ensure everyone has the same opportunity to participate in the digital economy. To date, we are the only credential service provider serving the public and private sectors that blends leading automated technologies with purpose-designed teams of human agents to ensure equitable access for all.

## The Benefits of AI and 1:1 Face Match

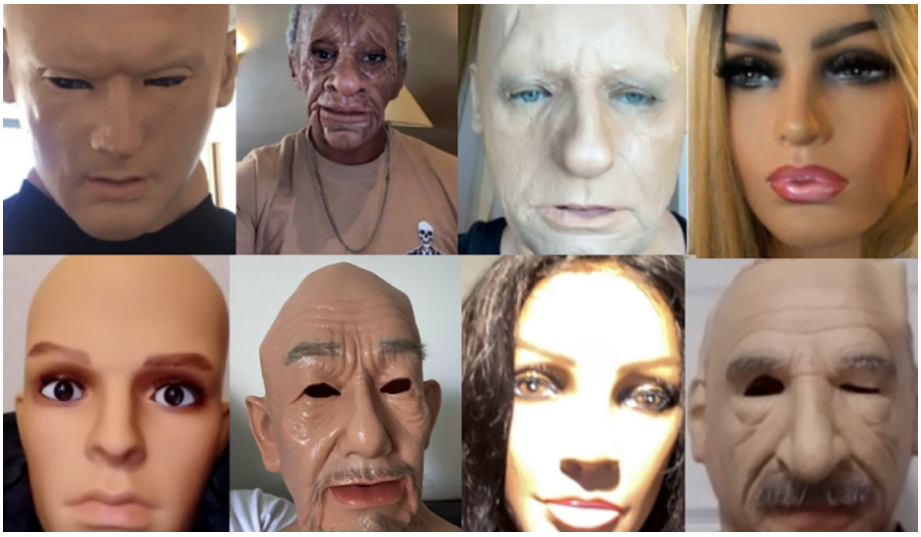
People and the government agencies that serve them see clear benefits when best-in-breed AI and 1:1 face-match technology are paired with humans-in-the-loop. Those benefits include:

- **A state workforce agency (SWA) can process genuine claims faster** – During the pandemic, the unemployment rate skyrocketed from 3.5% in February 2020 to 14.8% by April that year.[11] SWAs did not have the staff, technology infrastructure, or automated processes to keep up with the spike in demand. At the same time, international crime rings overwhelmed SWAs with fraudulent claims using stolen identity data, making it nearly impossible to distinguish legitimate applicants from fraudulent claims and causing large claim backlogs. By implementing NIST IAL2/AAL2 and PAD, SWAs stopped fraudulent claims while continuing to process legitimate claims the same day claimants completed identity verification. When Arizona implemented IAL2, new Pandemic Unemployment Assistance claims fell by 98.8% and existing claims fell by 68.3%, nearly all of which was fraud.[12] That enabled the SWA to focus on the smaller pool of legitimate claims and serve those applicants faster while dramatically reducing claim backlog.

11 Bureau of Labor Statistics, accessed December 2021. https://www.bls.gov/cps/

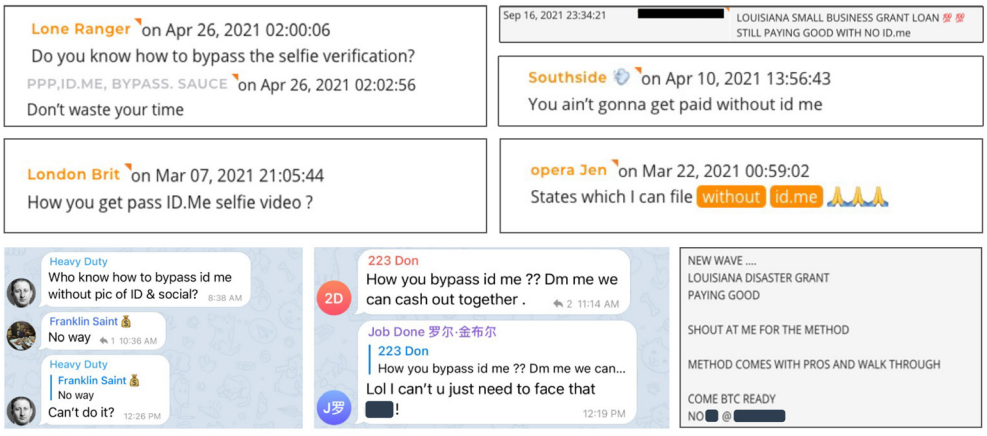
12 "How a Public-Private Partnership Provided Benefits to Eligible Individuals and Saved Billions for One State," September 2021. https://cdn.fedscoop.com/IDme-Arizona-PUA-Case-Study.pdf

- **Face liveness actively stops identity theft at scale** – Face liveness prevents an attacker from committing identity theft by ensuring the selfie submitted is an actual face and not an image, video or mask. That control actively blocked tens of thousands of identity theft attempts that would likely have succeeded and resulted in traumatized victims and lost funds.



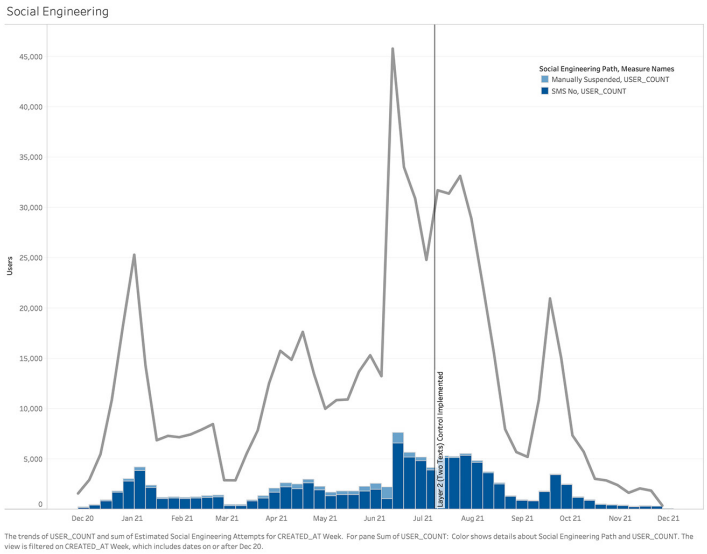
**PAD stopped criminals using masks from defrauding agencies.**

- **Face liveness deters bad actors from attempting fraud** – By the time fraud is detected, the criminal is typically long gone, leaving the government and the identity theft victim with no way to know who filed the fraudulent application. With PAD, the fraudster's selfie is preserved as part of the application audit trail. Thieves are understandably reluctant to provide an image of their true face when committing a crime, so face liveness helps deter fraud and the associated increase in case backlog. If a thief is brazen enough to submit a selfie while committing fraud, that information can be helpful to government agencies as they seek to recover stolen funds. Dark-web chatter among fraudsters underscores the strength of face liveness as a deterrent:

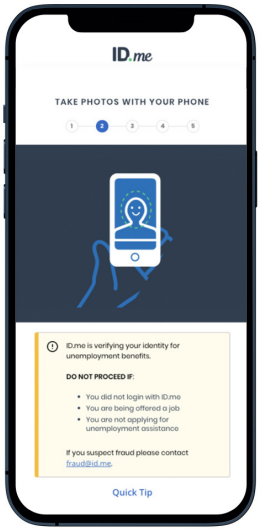



**Dark-web chatter illustrates face liveness effectiveness.**

- **Interstitial screens before the selfie stop imposter scams** – Social engineering involves identity thieves manipulating a victim into divulging confidential information or taking actions that compromise portions of the victim's identity. For example, people might be fooled into giving their Social Security number over the phone to a fraudster posing as a law enforcement agent or persuaded to upload a picture of a driver's license to a fake online job site. Currently, the most popular imposter scams involve lottery winnings, dating sites, job applications, and fake retail sites. The figures below show the scale and speed at which organized crime syndicates can ramp up social-engineering attacks, the interstitial screens ID.me used to pierce the scams, and examples of actual text exchanges between potential victims and attackers after the contextual warnings alerted the victim to the scam.




**Anti-social engineering controls enabled ID.me to stop tens of thousands of imposter scams a week.**




**Interstitial screens prior to the selfie provide context to stop social engineering attacks.**


**Validated user feedback from an exchange with an attacker shows the screens are effective.**

The rapid drop-off in attempted attacks reflects the effectiveness of controls enabled by the screens tied to the 1:1 face-matching process. The "Do Not Proceed" bullets actively target the most common false pretenses attackers use and let the victims know they are being tricked. Those measures have helped more than 100,000 people stop, or mitigate, the effects of identity theft over the past year.

- **It helps bring justice for those who have been exploited** – In October 2021, the Pennsylvania attorney general brought charges against a caregiver at a facility for people with disabilities for "stealing personal information of several intellectually disabled people in his care to fraudulently apply for and receive Pandemic Unemployment Assistance funding."[13] The caregiver was unable to navigate the ID.me unsupervised flow with the victim's information because of adherence to NIST IAL2 fraud controls. When the victim's application was escalated to a video chat with an ID.me Trusted Referee, it became clear the victim was being manipulated by the caregiver. The ID.me Trusted Referee alerted the state, and the Pennsylvania attorney general's office took swift legal action.
- **Makes account recovery easier if someone loses a device** – NIST SP 800-63B provides guidance on how to establish a new authenticator should someone lose access to a primary authenticator (typically a cellphone). Individuals can repeat the identity-proofing process or use an "abbreviated proofing process confirming the binding of the claimant to previously-supplied evidence."[14] The abbreviated pathway can be completed by submitting a selfie that is checked against the photo on the piece of identity evidence that was captured during original proofing. With a 1:1 face match between the new selfie and the image submitted during enrollment, account recovery can happen in seconds. In that scenario, people who lose their phones can recover their accounts and link to a new phone with less time tax than with other options, such as calling a help desk or going somewhere in person.
- **Prevents account takeovers** – Face liveness has prevented a significant number of account takeovers, particularly attacks against elderly users, because the criminal who stole the login credentials from the user is unable to pass the selfie check that ensures the same user who enrolled still controls the login.

13 State of Pennsylvania, Attorney General Josh Shapiro website, Accessed on November 16, 2021. https://www.attorneygeneral.gov/taking-action/press-releases/ag-shapiro-arrests-caregiver-for-stealing-intellectually-disabled-clients-personal-information-to-get-nearly-90k-in-unemployment-benefits/

14 National Institute of Standards and Technology Special Publication 800-63B, Section 6.1.2.3, accessed December 2021. https://pages.nist.gov/800-63-3/sp800-63b.html

# Crafting a Path to a More Equitable Society

***ID.me performed this analysis to improve understanding of AI and facial recognition as applied to authentication for government services. We hope it raises awareness of the benefits of AI in enabling faster service delivery, particularly during a time of crisis and heightened need. Using AI and biometrics in an ethical manner unlocks the convenience and time savings that are driving the growth of the digital economy in an equitable manner that includes all groups. Those gains should be realized with appropriate oversight.***

Currently, OMB M-19-17 directs agencies to use "Federally or commercially provided shared services…to deliver identity assurance and authentication services to the public." OMB also calls for a certification program to ensure those shared services meet the NIST 800-63-3 standards at a given assurance level. To further strengthen identity assurance services, policymakers could require human-driven relief valves and develop equity and inclusion metrics as part of that certification program.

On August 27, 2020, Pew reported, "States that were generous and quick to help workers were also quick to be targeted by scammers. In response, states have had to slow down the processing of claims, delaying payouts to people supposed to be getting them."

During the pandemic, agencies were unable to simultaneously scale services to meet demand while also effectively stopping fraud using manual processes. A hybrid approach that fused best in class AI algorithms with human reviewers as a relief valve proved vital. Getting this approach right is critical to scale services during a time of great need.

The best-performing systems are resilient. Combining algorithms with multiple layers of human review mitigates any potential bias that might arise.

**That approach offers the best path to equity and access for all.**

# Appendix A
# ID.me March 2021 Face-Match and Skin-Tone Analysis

In a March 2021 test, ID.me found that face match had a 95% success rate in practice. We picked the Social Security Administration for analysis as a broadly representative government agency that is not a target for fraud like state unemployment agencies. We then pulled a randomized sample of 627 individuals from the 5% of users who had failed face match. There are several key variables beyond the face-match algorithm that influence pass rates:

1. Document watermark refers to security features over the photo on the government ID that obscure the face and/or produce glare
2. "No apparent reason" represents a false negative when the face and the photo on the government ID appear to be a match with no apparent quality issues
3. "Reference ID photo low quality" refers to a poor-quality image of the government ID photo
4. Selfie face obscured
5. Selfie photo low quality
6. "Face change" means the selfie and the government ID could be the same person, but some **states issue licenses that are valid for decades**[15], so the reference photo is out of date.

Once we classified failure reasons, we used the Fitzpatrick Skin Phototype Classification (FSPC) framework to classify individuals: 1 being the lightest and 6 being the darkest. We ran a regression to see if there was a relationship between skin type and failure reason. As you can see from the P values on the following page, there is no statistically significant relationship between skin type and failure reason. No P value less than .05[15].

15 https://www.abc15.com/news/roads/why-arizona-driver-licenses-don-t-expire-for-decades

| | coef | sdd err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 2.0968 | 0.275 | 7.623 | 0.000 | 1.557 | 2.637 |
| **failure_reason[T. Document Watermark]** | 0.4894 | 0.396 | 1.237 | 0.217 | −0.288 | 1.266 |
| **failure_reason[T. No Apparent Reason]** | 0.4556 | 0.313 | 1.455 | 0.146 | −0.159 | 1.070 |
| **failure_reason[T.Reference ID Photo Low Quality]** | 0.3271 | 0.297 | 1.100 | 0.272 | −0.257 | 0.911 |
| **failure_reason[T.Selfie Face Obscured]** | 0.3860 | 0.341 | 1.133 | 0.258 | −0.283 | 1.055 |
| **failure_reason[T.Selfie Photo Low Quality]** | 0.2832 | 0.350 | 0.809 | 0.419 | −0.404 | 0.971 |
| **failure_reason[T.True Mismatch – Face Changed]** | 0.0503 | 0.299 | 0.168 | 0.867 | −0.537 | 0.638 |

**No statistical relationship between skin tone and face-match failure.**

We also ran a Chi-Square test for categorical variables and proportion tests for significant differences in proportions for group and reason while controlling for sample size. Neither test presented evidence of a relationship between skin tone and failure reason. Keep in mind that any individual who isn't matched in an unsupervised flow is not blocked from access and can still verify through supervised remote (video chat).

# Appendix B
# ID.me December 2021 Face-Match and Skin-Tone Analysis

In December 2021, ID.me tested for bias related to face match and skin tone per the Fitzpatrick Scale. We picked the IRS as a separate agency that is broadly representative and not an extreme target for fraud outside of tax season. We used 15,468 labeled images, collected in two sample sets for separate tests. Those samples were taken in two randomly sampled batches of users attempting to verify between November 17 and December 6. The first test run evaluated the selfie 1:1 match responses for correlation between the Fitzpatrick Scale number and the rates of selfie-match 1:1 failures for one sample set. The pass rate for all skin types in the first test averaged 98.5%.

The second sample set was used to run the same tests on liveness data. The pass rate for all skin types in the second test averaged 96.1%.

Variation in pass rates across skin types was measured in tenths of a percent and was not statistically significant. Both samples passed a Chi Square test for independence, indicating selfie-match and liveness failure rates were independent of skin type value on the Fitzpatrick Scale (p-value of 0.69 for liveness match and 0.60 for selfie match). We also ran regressions on the two sample sets to quantify any correlation between the Fitzpatrick Scale and selfie/liveness passes. Both coefficients of correlation were near zero (-.0013 for liveness match, and .000013 for selfie match).

# Appendix C
# ISO 30107–3 Conformance Test Results for iProov

Between July and August 2020, iBeta conducted a conformance test for iProov's PAD technology. iBeta is nationally accredited as a test lab to the requirements of ISO/IEC 17025:2017 by the National Voluntary Lab Accreditation Program (NVLAP). In 2011, iBeta was accredited by NIST under NVLAP for biometric testing under NIST handbook 150-25 and is considered an expert in the field of biometrics. In addition, iBeta procedures against the ISO 30107-3 PAD standard were audited by its accrediting body and iBeta's scope of accreditation was increased to include conformance testing to the ISO 30107-3 standard in April 2018.

During the tests, there were 12 species of presentation attacks (PAs) attempted against iProov capabilities 550 times across 10 test subjects. The tests were completed in two phases. In the first, testing was conducted in accordance with the contract for a level of spoofing technique that used only simple, readily available methods to create an artifact of the genuine biometric for use in the presentation attack, in other words, basic spoofs of an individual's face, such as with a printed picture. In the second, testing was conducted at a level of spoofing technique that used materials available for less than $300 and could be created in less than 24 hours, such as a high-quality resin mask. Level two tests represented a more sophisticated level of presentation attack than level one.

**According to the study:**

"The overall combined Imposter Attack Presentation Match Rate (IAPMR) for the system equates to an overall PA success rate of 0%. The False Non-Match Rate (FNMR) for the genuine subjects presenting was 0%, meaning that 0% of the presentations from a genuine subject were not accepted as either alive or authorized."