The Identity and Data Sciences Laboratory (IDSL)

at the Maryland Test Facility
1221 Caraway Ct. Suite 1070
Upper Marlboro, MD 20774

RE: Senate Bill 762 - Informational Only

March 8, 2022

On behalf of the Identity and Data Sciences Laboratory (IDSL) we are pleased to submit written informational testimony regarding Senate Bill 762 / House Bill 1046. We were also able to review and would like to comment regarding the amendment to be offered in the Judicial Proceedings Committee by Senator Sydnor. The IDSL is an independent research organization within Science Applications International Corporation (SAIC), specializing in independent test and evaluation of commercial biometric systems, including face recognition systems.

Since 2014, we have tested dozens of commercial biometric technologies in various government use-cases[1]. Our research shows that, when commercial face recognition systems are used to establish the identity of individuals, they can make errors, sometimes conflicting with notions of 'fairness' or 'equitability'. While top performing systems can work well across demographic groups, our experience suggests that vendor-reported efficacy claims may not always align with real-world performance. There is also significant variation in performance across vendors.

Face recognition systems are complex and international standards define several types of biometric testing including technology testing, scenario testing, and operational testing. The different tiers of testing are needed because, in addition to the matching algorithm, performance of these systems depends on implementation details. These include gallery size, quality of the face photos used for matching, the demographics of the individuals in the photos, as well as the training of human reviewers of system results.

Maryland houses significant expertise in testing biometric systems. For example, the National Institute of Standards and Technology (NIST) in Gaithersburg runs the Face Recognition Vendor Test (FRVT) which performs technology testing of algorithms in isolation. The IDSL in Upper Marlboro has performed a variety of tests on behalf of the Department of Homeland Security Science and Technology Directorate (DHS S&T), but specializes in scenario testing, which tests full systems in a simulated environment. Both NIST and the IDSL have successfully integrated commercial systems into test infrastructure by asking vendors to implement a standardized API [**Amendment No. 2 (B) (1)**], to measure performance. No one type of testing is sufficient in isolation, however, our experience suggests the following approach [**2-506 (A) (3)**]: (1) Pick initial vendors based on NIST algorithm testing; (2) test vendor performance with scenario testing using operationally relevant images, galleries, and demographics (e.g., probes and reference galleries that reflect the sizes and demographics of those in Maryland's intended operational use); and (3) use test results to select a final vendor. Additionally, face recognition algorithms are updated frequently; once, purchased

[1] The IDSL staffs DHS S&T's Maryland Test Facility (MdTF). https://mdtf.org

those selecting the algorithm should validate if the specific version tested matches the version being purchased.

Testing of biometric systems requires a significant quantity of data. Test data may be gathered from new volunteers in a scenario test (typically hundreds of volunteers are needed for a statistically significant evaluation), or use previously acquired photos linked with ground truth self-reported demographic information and independently measured skin tone [**Amendment No. 2 (A) (1)**]. Web-scraped data are generally inappropriate as they are not linked with ground-truth demographic information. Ideally, data for testing should be acquired with informed consent as well as privacy protections. There are few appropriately labeled, responsibly collected, datasets of sufficient size to test modern face recognition systems along the subpopulations delineated in the ammendment [**Amendment No. 2 (A)(1)**].

For these reasons, test datasets must remain sequestered. If technology developers have access to test datasets, they may use them in the creation of their algorithms. This will lead to good performance on tests for trivial reasons, like knowing the questions and answers on a test ahead of time. For this reason, sharing test data with technology developers is not considered good practice [**Amendment No. 2 (B) (3)**]. If test data is shared, subsequent testing would require all new data. If new operational data cannot be shared [**Amendment No. 2 (D)**], then these new data would have to be collected specially, which may carry significant costs.

Though international standards for testing biometric systems have existed since 2006, there is currently no standard methodology for testing biometric systems for fairness. Our group has examined these issues in great technical detail. Indeed, two of the authors of this letter (J.J.H. and Y.B.S.) are co-editors of a draft international standard on measuring demographic differentials in biometric system performance. There are many fairness metrics proposed for evaluating biometric systems, most of which include mathematical differences and ratios. Picking the right metric is extremely important to understand the result. For example, one can obtain large ratios of error rates observed between two groups (e.g., 10-100 times) even though differences between error rates are vanishingly small. More important still, there is no standard statistical criteria for determining what constitutes unfair difference in system performance. This criterion for what constitutes a "material" difference cannot come from a statistical or mathematical formula, it must be developed by policy [**Amendment No. 2 (B) (2)**].

Testing face recognition systems is needed to select appropriate commercial technologies and to ensure they work well within a specific use-case. We have provided similar information to a recent RFI from the White House Office of Science and Technology Policy (OSTP; attached). We hope this testimony will inform further development of this important legislation.

Very Respectfully,

On behalf of the Identity and Data Sciences Laboratory

Jerry L. Tipton, Executive Director, jtipton@idslabs.org
Yevgeniy B. Sirotin, PhD, Technical Director, ysirotin@idslabs.org
John J. Howard, PhD, Lead Data Scientist, jhoward@idslabs.org

ATTACHMENT:

IDSL Response to the White House Office of Science and Technology Policy RFI Document No: 2021-21975

1.0 About the Identity and Data Sciences Laboratory (IDSL)

The Identity and Data Sciences Laboratory (IDSL) is an independent research organization within SAIC, a technology integrator for the US government. The IDSL is comprised of scientists, engineers, IT specialists, and program managers with demonstrated expertise in the test and evaluation of AI systems.

Since inception, the IDSL has carried out authoritative analyses and reporting on the performance of biometric identity systems, including face recognition systems. Much of our work has been in support of the Department of Homeland Security Science and Technology Directorate (DHS S&T). The IDSL operates the Maryland Test Facility (MdTF) in support of research conducted on behalf of the DHS S&T Biometric and Identity Technology Center (BI-TC). Starting with our work on the Air Entry-Exit Re-engineering (AEER) project we have tested well over 200 commercial biometric technologies in varied use-cases. Our technology evaluations have been provided to inform government agencies (DHS S&T, CBP, TSA, USCIS, OBIM, DOD, DOJ, and others) as well as published in peer-reviewed scientific journals[2]. Our expert staff are regularly invited to present our findings at conferences within the US and internationally. IDSL applied research addresses topics including biometric system performance, demographic group fairness, and human-algorithm teaming. We are using this insight to inform the development of international standards, including technical editorship of ISO/IEC 19795-10 on quantifying biometric system performance variation across demographic groups.

Given this relevant background, we are pleased to respond to the White House Office of Science and Technology Policy (OSTP) request for information (RFI) titled "Notice of Request for Information (RFI) on Public and Private Sector Uses of Biometric Technologies". In the sections below, we provide responses to topic areas outlined within the RFI.

2.0 Responses to RFI Topic Areas

2.1 Descriptions of use of biometric information for recognition and inference

As defined by OSTP, the definition of biometric technology to include both individual recognition and cognitive/emotional state inference encompasses a wide range of disparate technology. Because of foundational differences in these two kinds of computer applications, care is often taken to separate the two in the scientific community. For example, there are internationally adopted standards that define the term "biometrics" as "automated recognition of individuals" based on their behavioral and biological characteristics" (emphasis ours)[3]. This definition has also previously been adopted by agencies in the U.S. Government[4]. By this definition, biometric recognition involves a comparison between two biometric samples to determine whether they are of the same individual.

[2] MdTF Publications. https://mdtf.org/Research/Publications.

[3] ISO/IEC 2382-37:2017 Information technology — Vocabulary — Part 37:"Biometrics Recognition" term 37.01.03. https://www.iso.org/standard/66693.html

[4] DHS OBIM defines a biometric as "a measurable biological (anatomical and physiological) and behavioral characteristic that can be used for automated recognition". https://www.dhs.gov/biometrics

Biometric recognition has well defined scientific underpinnings, metrics, and international standards that have been in existence for nearly 20 years[5]. Indeed, biometric systems may be one of the most well tested current applications of artificial intelligence (AI)[6]. For nearly a decade, biometric systems have been deployed in a variety of scenarios including to facilitate identity determination at international borders and airport checkpoints, for individual identification in both public and commercial settings including the identification of missing persons and those involved in human trafficking, and for access to personal electronic devices.

In contrast, technology for inference of cognitive and/or emotional states based on a single sample are varied in their domain of application and poorly understood. The scientific basis for these technologies also varies dramatically (some basis for emotion recognition[7] vs no basis for criminality[8]). Additionally, we are not aware of any international standards for the test and evaluation of these systems. Despite growing commercial deployment in areas such as hiring and exam monitoring, these technologies are rarely, if ever, vetted for validity by independent third parties.

As an entity specifically focused on AI system test and evaluation, the bulk of our responses to this RFI are centered on biometric technology as used for recognition since this is where our primary experience lies. Our position is that it may be timely to consider similar scrutiny to other AI systems in the public domain.

2.2 Procedures for and results of data-driven and scientific validation of biometric technologies

With support from the Department of Homeland Security Science and Technology Directorate, the IDSL conducts data-driven scientific evaluations of biometric technology in government use-cases. At a high level, there are three kinds of biometric evaluations as defined by ISO standards[9] enumerated below. In Sections 2.2.1 – 2.2.3, we outline each evaluation type, including measurement setup, evaluation procedure, specific measures, outcomes and error rates.

Technology evaluations are typically centered on a specific component of a biometric system (e.g. a matching algorithm) and use previously acquired biometric datasets with large sample sizes. This type of testing is appropriate for measuring the limits of a technology's performance and for comparison of different technologies. This testing is not appropriate for answering questions about how a technology performs in a specific application.

[5] ISO/IEC JTC 1/SC 37 Biometrics. https://www.iso.org/committee/313770.html

[6] NIST: Biometrics. https://www.nist.gov/programs-projects/biometrics.
DHS S&T Biometric and Identity Technology Center (BI-TC). https://www.dhs.gov/science-and-technology/BI-TC.
The Maryland Test Facility. https://mdtf.org.

[7] Barrett, Lisa Feldman, et al. "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements." Psychological science in the public interest 20.1 (2019): 1-68.

[8] Bowyer, Kevin W., et al. "The "Criminality From Face" Illusion." IEEE TTS 1.4 (2020): 175-183.

[9] ISO IEC 19795-1: Information technology–biometric performance testing and reporting-part 1: Principles and framework. https://www.iso.org/standard/73515.html.

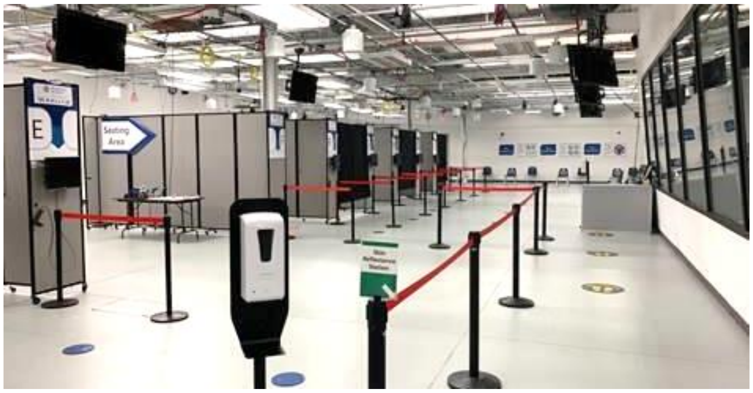

Figure 1. Maryland Test Facility test bay set up for a "Rally" scenario test.

Scenario evaluations center around a specific technology use-case (e.g. airplane boarding) and test a full multi-component biometric system (i.e. including any acquisition devices, databases, and algorithms) with test volunteers in a controlled environment. This type of testing gathers new biometric samples to answer questions about how a technology performs for a specific intended use (**Figure 1**).

Operational evaluations assess the performance of a technology in the fielded environment. This testing measures the performance of the system within a specific location and environment (e.g. a face recognition system installed in at a specific airport terminal). While most operationally relevant, reduced experimental control in operational evaluations makes it harder to identify the key factors influencing performance.

2.2.1 Technology evaluations

By far the most common category of biometric evaluation are what's known as technology evaluations. Technology evaluations typically rely on large static test datasets and can be used to test performance limits and track the performance of algorithms over time, motivating innovation. Tests are typically executed on biometric algorithms in isolation, disentangling them from the larger workflows of full operational biometric systems (i.e. cameras, databases, administrative systems, etc.).

The IDSL regularly executes technology evaluations to report on both the state of the biometric industry and industry progress. To execute technology evaluations, the IDSL maintains a sophisticated data storage, processing and reporting infrastructure in house at the Maryland Test Facility. This computational testbed consists of over 25 distinct server systems, 100 virtualized software platforms for redundancy, and 20 TB of on-premise storage.

The protocols, measures, and outcomes for technology testing are defined in the international standard ISO/IEC 19795-2, which has been in place since 2007[10]. Typically, experimental setup in a technology test involves a large static dataset of biometric samples with ground truth. Biometric algorithms are used to create biometric templates, or mathematical models of the physiological sample. These templates can then be compared to calculate a similarity score. Once this process has been executed on many biometric sample

[10] ISO/IEC 19795-2:2007 Information technology — Biometric performance testing and reporting — Part 2: Testing methodologies for technology and scenario evaluation. https://www.iso.org/standard/41448.html

pairs (face pairs, iris pairs, etc.) the generated scores are separated into two categories; those that came from biometric samples that should match (individual A's face image on day 1 and individual A's face on day 2) and those that should not (individual A's face and individual B's face). These pairs are called mated and non-mated pairs respectively. Using these pairs, two foundational error rates for a biometric algorithm can be calculated, namely the false non-match rate and the false match rate. Both these error rates measures are specific to a match or discrimination threshold. Its common in technology testing for these error rates to be calculated over a range of thresholds to produce summary statistics, such as detection error tradeoff curves.

The main benefits of technology evaluations of biometric systems lie in their reproducibility. This is advantageous because 1) the findings can be replicated by others and used to improve their systems (assuming data availability) and 2) the findings can be replicated longitudinally as algorithms or other system components improve. In this way technology evaluators can monitor and report on industry progress. We have previously used technology evaluations to identify a phenomenon named "demographic clustering", by which face recognition algorithms tend to score different people of the same race, age, and gender as more similar than those who do not share demographic characteristics[11]. We first pointed out this "homogeneity effect" in 2019 and subsequently replicated it with numerous algorithms and on other datasets[12].

Technology testing has important limitations. Much like comparing two formula 1 race cars on a test track, you are able to see what is achievable, but you are unlikely to see comparable performance driving your sedan around town. Technology testing will miss important aspects of operational system performance. For example, a technology evaluation may not discover a scenario in which a facial recognition camera systematically cannot find faces (and therefore take pictures) of individuals with darker skin, since these evaluations starting point is captured images. Furthermore, the static nature of the datasets used in technology evaluations means that they often do not represent changing circumstances in the real world. For example, when the COVID-19 pandemic led to large scale public masking requirements, the datasets used in typical face recognition technology evaluations no longer reflected the facial characteristics of individuals a face recognition system was likely to encounter in situations like an airport or border crossing.

In summary, technology evaluations of biometric technologies are well defined processes that provide important information, particularly to biometric system developers. However, they are not sufficient to anticipate the full range of issues a biometric system might experience once deployed in a robust, operational environment. They are one part of a larger, necessary testing regime to ensure the effectiveness and equitability of biometric systems.

[11] Howard, Sirotin, Tipton, Vemury. Quantifying the Extent to Which Race and Gender Features Determine Identity in Commercial Face Recognition Algorithms. DHS S&T Technical Paper Series. (2021). https://www.dhs.gov/sites/default/files/publications/21_0922_st_quantifying-commercial-face-recognition-gender-and-race_updated.pdf

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

2.2.2 Scenario evaluations

Scenario evaluations of biometric technologies simulate a full biometric application and its real-world deployment environment. Unlike technology evaluations, scenario evaluations measure error and success rates on full biometric systems (i.e., algorithms, acquisition devices like cameras, and any needed databases). Further scenario evaluations measure performance using new data collected from test volunteers. In every new evaluation, volunteers utilize biometric systems just as they would in a real-world deployment, allowing unique insights into the efficiency of the system (e.g. how long it takes to use) and on human perceptions of the system (e.g. how satisfied are the users).

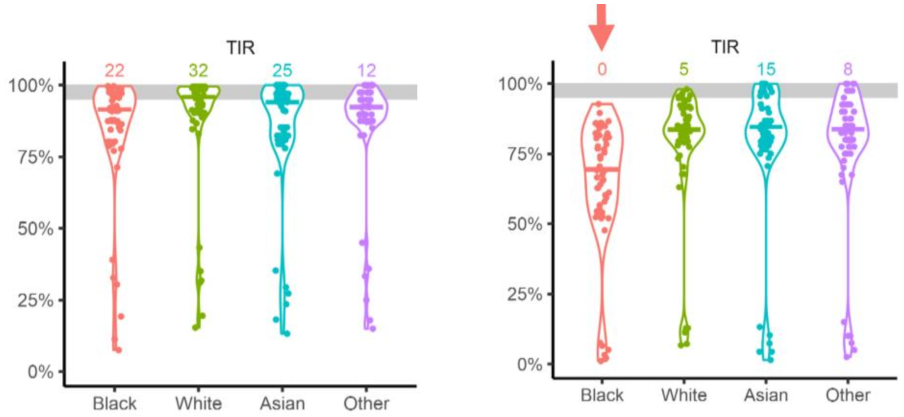

Figure 2. *True Identification Rate (TIR) of face recognition systems without face masks (left) and with face masks (right) disaggregated by self-identified Race. Note greater reduction (arrow) due to masks for those self-identifiyng as Black.*

To date, the IDSL has primarily focused on scenario evaluations of both staffed and automated biometric systems within the travel environment[13]. We curate and maintain an ethically collected structured dataset of over 137,000 face, fingerprint, and iris images of over 2,000 unique persons together with metadata on demographics and phenotypes (e.g. skin tone). For our scenario tests, we recruit volunteers from the local area stratified by race, gender, and age or other factors as needed for each evaluation. We have tested well over 200 face, fingerprint, and iris recognition systems with over 5,000 unique volunteer visits to the Maryland Test Facility. The IDSL uses dedicated data processing systems for computing standard measures of biometric performance and generating reports.

Using this scenario test model, scientists at the IDSL have identified important new insights into biometric performance. For example, in a widely cited 2018 study that explored the effect of camera on bias, we found evidence that differential performance in face recognition could largely be traced to differences in camera's abilities to capture high quality photographs of individuals with difference skin tones[14]. This impact of camera had largely been ignored in discussions of "bias" in face recognition but plays a key role in creating a more equitable system. Additionally, using the scenario test model the IDSL was able in 2020 to collect the first

[13] Howard, Blanchard, Sirotin, Hasselgren, Vemury An Investigation of High-Throughput Biometric Systems: Results of the 2018 Department of Homeland Security Biometric Technology Rally. https://mdtf.org/publications/rally-results.pdf.

[14] Cook, Howard, Sirotin, Tipton, Vemury. Demographic Effects in Facial Recognition and their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. https://mdtf.org/publications/demographic-effects-image-acquisition.pdf

dataset of masked individuals since the onset of the COVID-19 pandemic. We were able to quantify the expected reduction in face recognition performance due to masked face occlusion and critically demonstrated that this performance reduction was not equivalent across demographic groups (individual with darker skin saw larger reductions in performance than those with lighter skin, **FIGURE 2**)[15]. This insight motivated improvements in masked face recognition performance across industry and helped created more equitable face recognition systems.

Lastly, we have found that scenario testing at the IDSL forecasts error cases in the operational environment. In particular, scenario testing predicts the use errors and differences in performance associated with demographic factors. On the other hand, results observed in technology tests depend critically on the type of data used for the evaluation. For instance, the performance of face recognition technologies in NIST's FRVT tests depends critically on the type of dataset used[16]. In our own assessments, we find that the performance of system components is inter-dependent with algorithm results depending strongly on the acquisition camera used (**FIGURE 3**)[17]. We strongly believe that, like other forms of AI, biometric technologies must be proven in scenario tests in order to understand their likely performance within the operational environment.

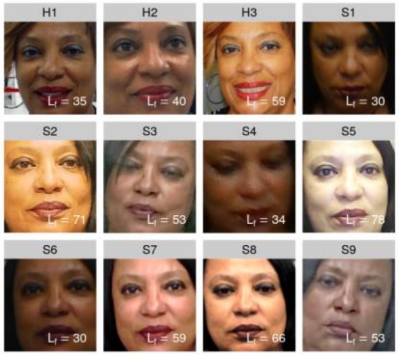

Figure 3. *Images of a person gathered using different biometric cameras. Note the change in appearance and skin tone. Images S1-S9 were collected on the same day under consistent lighting conditions.*

2.2.3 Operational evaluations

The final variety of biometric system evaluation is known as an operational evaluation. The protocols and procedures for this form of testing is defined in international standard ISO/IEC 19795-6, which was published in 2012[18]. Operational evaluations provide the most direct insight into how a biometric system is performing as deployed in a given implementation. However, despite their value, operational evaluations of biometric

[15] Y. B. Sirotin and A. R. Vemury. "Demographic variation in the performance of biometric systems: Insights gained from large-scale scenario testing." In Virtual Events Series – Demographic fairness in biometric systems. EAB, March 2021. https://mdtf.org/publications/ EAB2021-Demographics.pdf.

[16] Grother, Patrick, et al., "Onoging Face Recognition Vendor Test (FRVT) Part 1: Verification."

[17] Hasselgren, Jacob A., et al., "A scenario evaluation of high-throughput face biometric systems: select results from the 2019 Department of Homeland Security Biometric Technology Rally." DHS S&T Technical Paper Series. (2020). https://www.dhs.gov/sites/default/files/publications/2021_st-01_2019selectrallyresultstip20201104_revised_3046.pdf

[18] ISO/IEC 19795-6:2012 Information technology — Biometric performance testing and reporting — Part 6: Testing methodologies for operational evaluation. https://www.iso.org/standard/50873.html

systems can be challenging to resource and execute properly. Consequently, they are relatively rare compared to scenario and laboratory evaluations of biometric systems. The two main challenges when conducting operational evaluations of biometric systems are lack of experimental control and lack of ground truth information. For example, it can be arduous to collect accurate race, gender and age information from people in crowded operational environments, like airports or train stations. It can also be challenging attributing observed effects directly to specific causes because of many nuicanse factors.

To perform operational evaluations, the IDSL team goes on location to observe and record the operational environment, the technology configuration, and first-hand observations of user interactions with the system. The IDSL can receive and process operational sample-based and transactional data to generate performance measures. We believe operational evaluations of biometric systems provide the most direct evidence of system performance in the field to inform system developers and system owners.

2.3 Security considerations associated with a particular biometric technology

Discussion topic 3 in OSTP's RFI deals with the security of biometric systems, particularly around spoofing and more traditional software system security (i.e. encryption, data access/audit, etc.). We anticipate many respondents will provide material on these two topics. However, we wanted to raise a security issue that OSTP might not yet be aware of that relates specifically to face recognition applications. Often when face recognition is used for security applications, the digital images that require identification can come from poor quality cameras and challenging environments. There is a strong incentive to improve the utility of such low-quality images for biometrics, especially when this may help solve a crime.

However, the performance of biometric systems with altered digital images, even if altered with the intent to enhance, is generally not well understood and has been suggested to lead to potential law enforcement errors[19]. Further, recent advances in AI have made it easier to perform such alterations without needing technical skill[20]. This creates additional concerns regarding privacy whereby security equipment previously suitable only for detecting suspicious activity may now become useful for biometric surveillance.

To avoid errors and privacy implications that may be caused by image manipulation in security applications, it is important that biometric systems include specific descriptions of their intended context of use and that any performance information be clearly associated with this context of use.

2.4 Exhibited and potential harms of face recognition technology

The deployment of face recognition technologies undoubtedly carries with it potential harms, some of which have been realized as these technologies are increasingly used in the real world. First, in regards to the validity of the science, there is little doubt the human face contains characteristics that allow for individual recognition. Human beings innately perform such functions on a daily basis when we recognize friends,

[19] Garvie, Clare, et al., "The perpetual line-up. Unregulated police face recognition in America". Georgetown Law Center on Privacy & Technology. (2016). https://www.perpetuallineup.org/

[20] Some examples: research from Google (https://ai.googleblog.com/2021/07/high-fidelity-image-generation-using.html) and of a tool easily available online (https://github.com/TencentARC/GFPGAN).

family, co-workers, etc. It stands to reason that computer processes could similarly carry out such tasks, a notion which has been repeatedly validated by over 20 years of government and industry testing.

However, just because a given technology works in the general case, does not mean it works equally well for all groups of people. Additionally, a technology that works well in the general case can also have idiosyncrasies that cause it to fail in predictable ways. Both of these conditions are true for face recognition. Many scientists, IDSL staff included, have documented error rates that can differ for individuals based on their demographics in face recognition. We coined the, now widely adopted, term "demographic differentials" to describe these effects in 2018[21]. While studying these phenomena is important, IDSL scientists have also pointed out that solving for this situation may not fully solve issues of "bias" in face recognition. In 2021, IDSL scientists highlighted an often overlooked but nearly universal characteristic of face recognition. Face recognition algorithms judge different individuals who share demographic characteristics (same race, gender age, etc.) as more alike than those that don't. We used the term "broad homogeneity" to describe this effect and pointed out that no other major biometric modality does this, yet it has somehow become accepted in face recognition[22].

We believe this clustering by demographics may be one source of potential harm in face recognition deployments when used for law enforcement. The fact that broad homogeneity exists means that identifications against galleries that are demographically skewed (majority male, for example) could have unequal false positive identification rates. Implementers of face recognition workflows should be aware of this effect and its consequences. Training may help avoid adverse impacts that stem from this phenomenon.

2.5 Exhibited and potential benefits of face recognition technology

The deployment of face recognition systems has undoubtedly benefitted the general public in many ways. One of the clearest examples is in the travel environment, where face recognition applications have sped airplane boarding and border crossing. Prior to the introduction of automated face recognition in these environments, identity verification tasks were performed by exclusively by humans. However, humans have well documented shortcomings when it comes to identifying unfamiliar faces. Humans also have limitations in terms of attention. This makes automated face recognition an attractive choice to both improve the effectiveness and efficiency in these environments.

2.6 Governance programs, practices, and procedures

All IDSL scenario test activities conducted at the Maryland Test Facility receive approval from an external Institutional Review Board (IRB) to ensure that ethical and data safeguards are met. Additionally, all data collected as part of our work with DHS S&T is maintained in accordance with a Privacy Threshold Analysis approved by the DHS Privacy Office. As part of standard practices required by the IRB, all human-subjects that participate are properly informed about the test and provide explicit consent to participate.

[21] Howard, Sirotin, Vemury. The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance. https://mdtf.org/publications/broad-and-specific-homogeneity.pdf
[22] Ibid., 10

The IDSL conducts two forms of informed consent for all test events: group consent and individual consent. In the group consent, all human-subjects are informed of what data will be collected and how their data will be protected. In the individual consent, human-subjects are called into private interview rooms with doors and white noise to guarantee privacy to each human-subject while going over consent forms. Each human-subject is asked for explicit permission to reproduce any images collected during the test in publication materials; subjects that opt-out are not excluded from the test.

All data collected by the IDSL is associated with a unique subject-ID, separated from any personal information. This protection is to avoid personally identifiable information (PII) from being leaked or compromised. The IDSL's datasets are also sequestered to prohibit datasets from being taken advantage of by developers of AI/ML systems. Developers of AI/ML systems will leverage all available information in developing their system, but this can result in 'overfitting' (a phenomenon where you can do better on the data you know and paradoxically worse on new data) or even cheating[23]. For this reason we limit access to our datasets and routinely gather new data to prevent such practices, even when unintentional.

We believe technology and scenario evaluations play a critical role in biometric system governance prior to system deployment by reducing the odds that non-performant or unfair systems are put into real-world applications. However, following system deployment, additional performance auditing steps are also necessary to ensure that real-world conditions have not adversely impacted the expected performance of a biometric system. Because post-deployment performance evaluations are likely to contain PII collected outside the lab context, the IDSL utilizes separate systems for processing data gathered as part of an operational evaluation. Operational data used for performance evaluations resides on Government systems granted Authority to Operate (ATO) and is used in accordance with any required Privacy Threshold Assessments. Steps and considerations when conducting post deployment, operational evaluations of biometric systems are discussed in Section 2.2.3.

3.0 The case for requiring independent testing of biometric systems

As real-world deployments of AI systems multiply, the public is becoming increasingly aware of the need to evaluate the performance of AI systems. Our research shows that, when these systems are used to establish the identity of individuals and make inferences about individuals, they can make errors, sometimes conflicting with notions of 'fairness' or 'equitability'. Our experience suggests that vendor-reported efficacy claims may not always align with real-world performance. Depending on the application, biometric system errors may carry significant costs or harms at both the individual and group level[24].

[23] Markoff, John. "Baidu team is barred from A.I. competition." The New York Times. (2015). https://www.nytimes.com/2015/06/04/technology/computer-scientists-are-astir-after-baidu-team-is-barred-from-ai-competition.html and Quach, K. (2020, June 18). How a kaggle grandmaster cheated in $25,000 ai contest with hidden code – and was fired from Dream SV job. The Register - Biting the hand that feeds IT. Retrieved January 14, 2022, from https://www.theregister.com/2020/01/21/ai_kaggle_contest_cheat/

[24] Hill, Kashmir. "Wrongfully accused by an algorithm." The New York Times (2020). https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Despite technology developers racing to create and implement AI systems, few entities have the capability and focus, like the IDSL, to test the performance of these systems. The situation is comparable to the field of drug development prior to The Federal Food, Drug, and Cosmetic Act of 1938, which required new drugs to be shown safe and prohibited false therapeutic claims[25]. AI systems may not have direct effects on human life, but their increasing ubiquity and scale also carry the potential for significant harms.

Some recent discussions have focused on AI audits as means to ensure that harms of AI systems are managed[26]. While important, we believe that audits in the absence of independent third-party *performance testing* are insufficient to ensure that systems meet required benchmarks for performance and equitability.

The IDSL has a unique mission to evaluate biometric systems to better understand their likely performance in the field and to provide quantitative empirical evidence to inform analyses of these systems' potential harms, including harms to protected demographic groups. Currently, there is little incentive for companies to perform independent third-party tests of their biometric technology products. Conversely, companies have strong incentives to present optimistic performance claims in marketing that conflate results of technology testing performed during AI training and real-world performance.

Without robust regulations and requirements for rigorous scientific testing, like the kind carried out by the IDSL, few biometric system developers have the incentive to test their systems. Indeed, the US government currently shoulders much of the cost associated with testing these technologies. The costs of deploying untested systems will be realized in unexpected technology failures, including potentially unfair systems. These issues may be realized only after deployment, when changes or adjustments become more costly. Worse still is the possibility that such issues may simply go undetected, leading to increasing opportunity periods for harms to manifest. This will undermine public trust in biometric systems.

We believe that independent third-party scenario and operational testing with demographically diverse people should be a prerequisite to marketing biometric systems for any high-risk applications that carry potential for harms at the individual or the group level. We hope the information we have provided herein can inform the development of an AI bill of rights[27].

Durkin, Erin. "New York tenants fight as landlords embrace facial recognition cameras." The Guardian (2019). https://www.theguardian.com/cities/2019/may/29/new-york-facial-recognition-cameras-apartment-complex

[25] FDA. "Milestones in U.S. Food and Drug Law." https://www.fda.gov/about-fda/fda-history/milestones-us-food-and-drug-law

[26] The New York City Council - File #: Int 1894-2020 (nyc.gov) https://legistar.council.nyc.gov/LegislationDetail.aspx

[27] Lander, Eric and Nelson, Alondra. "ICYMI: WIRED (Opinion): Americans Need a Bill of Rights for an AI-Powered World." The Office of Science and Technology Policy. (2021). https://www.whitehouse.gov/ostp/news-updates/2021/10/22/icymi-wired-opinion-americans-need-a-bill-of-rights-for-an-ai-powered-world/