Stanford University
Human-Centered
Artificial Intelligence

Evaluating Facial Recognition Technology: A Protocol for Performance Assessment in New Domains

A WHITE PAPER FOR STANFORD'S INSTITUTE FOR HUMAN-CENTERED ARTIFICIAL INTELLIGENCE

Daniel E. Ho
Emily Black
Maneesh Agrawala
Li Fei-Fei

DISCLAIMER

The Stanford Institute for Human-Centered Artificial Intelligence (HAI) is a nonpartisan research institute, representing a range of voices. The views expressed in this White Paper reflect the views of the authors. Workshop attendees are listed only for informational purposes only.

NOVEMBER 2020

Table of Contents

Contributors

PRINCIPAL AUTHORS

Daniel E. Ho, Stanford University
Emily Black, Carnegie Mellon University
Maneesh Agrawala, Stanford University
Li Fei-Fei, Stanford University

OTHER CONTRIBUTORS

Christopher Wan, Stanford University
Evani Radiya-Dixit, Stanford University
Ohad Fried, Stanford University

CO-SIGNATORIES

Elizabeth Adams, Stanford University
Stephen Caines, Stanford University
John Etchemendy, Stanford University
Iacopo Masi, University of Southern California
Erik Learned-Miller, University of Massachusetts, Amherst
Harry Wechsler, George Mason University

ACKNOWLEDGMENTS

We acknowledge the generosity of all of the workshop participants and colleagues, without whom this effort would never have materialized. We thank Russell Wald, Michael Sellitto and Nazila Alasti for leadership in organizing the workshop, Celia Clark for her exceptional help in running the virtual convening, Danielle Jablanski, Marisa Lowe, Ohad Fried, Evani Radiya-Dixit, and Chris Wan for serving as rapporteurs, and helpful comments from Joy Buolamwini, Amanda Coston, John Etchemendy, Tatsu Hashimoto, Peter Henderson, Erik Learned-Miller, Michael Sellitto, and Christine Tsang.

Preface

In May 2020, Stanford's Institute for Human-Centered Artificial Intelligence (HAI) convened a half-day workshop to address the question of facial recognition technology (FRT) performance in new domains. The workshop included leading computer scientists, legal scholars, and representatives from industry, government, and civil society (listed in the Appendix). Given the limited time, the goal of the workshop was circumscribed. It aimed to examine the question of operational performance of FRT in new domains. While participants brought many perspectives to the workshop, there was a relative consensus that (a) the wide range of emerging applications of FRT presented substantial uncertainty about performance of FRT in new domains, and (b) much more work was required to facilitate rigorous assessments of such performance. This White Paper is the result of the deliberation ensuing from the workshop.

FRT raises profound questions about the role of technology in society. The complex ethical and normative concerns about FRT's impact on privacy, speech, racial equity, and the power of the state are worthy of serious debate, but beyond the limited scope of this White Paper. Our primary objective here is to provide research- and scientifically-grounded recommendations for how to give context to calls for testing the operational accuracy of FRT. Framework legislation concerning the regulation of FRT has included general calls for evaluation, and we provide guidance for how to actually implement and realize it. That work cannot be done solely in the confines of an academic lab. It will require the involvement of all stakeholders — FRT vendors, FRT users, policymakers, journalists, and civil society organizations — to promote a more reliable understanding of FRT performance. Since the time of the workshop, numerous industry developers and vendors have called for a moratorium on government and/or police use of FRT. Given the questions around accuracy of the technology, we consider a pause to understand and study further the consequences of the technology to be prudent at this time.

Adhering to the protocol and recommendations herein will not end the intense scrutiny around FRT, nor should it. We welcome continued conversation around these important issues, particularly around the potential for these technologies to harm and disproportionately impact underrepresented communities. Our limited goals are to make concrete a general requirement that appears in nearly every proposed legislation to regulate FRT: whether it works as billed.

We hope that grounding our understanding of the operational and human impacts of this emerging technology will inform the wider debate on the future use of FRT, and whether or not it is ready for societal deployment.

1. Introduction

Facial recognition technology (FRT), namely the set of computer vision techniques to identify individuals from images, has proliferated throughout society. Individuals use FRT to unlock smartphones,[1] computer appliances,[2] and cars.[3] Retailers use FRT to monitor stores for shoplifters and perform more targeted advertising.[4] Banks use FRT as an identification mechanism at ATMs.[5] Airports and airlines use FRT to identify travelers.[6]

FRT technology has been used in a range of contexts, including high-stakes situations where the output of the software can lead to substantial effects on a person's life: being detained overnight at an airport[7] or being falsely accused of a crime, as was the case for Robert Williams and Michael Oliver.[8] A 2016 study reports that one out of two Americans are involved in a "perpetual line-up" (i.e., an ongoing virtual police lineup), since local and federal law enforcement regularly perform facial recognition-based searches on their databases to aid in ongoing investigations.[9] Beyond the effects of current use of FRT, widening the deployment of FRT to continuous surveillance of the public has the potential to change our use of public spaces,[10] our expectations of privacy, our sense of dignity, and the right to assemble.[11]

The widespread use of FRT in high-stakes contexts has led to a loud call to regulate the technology — not only from civil society organizations,[12] but also by the creators and vendors of FRT themselves. IBM, for instance, has discontinued its sale of "general purpose facial recognition software," stating that "now is the time to begin a national dialogue on whether and how facial recognition technology should be employed by domestic law enforcement agencies," offering to work with Congress to this end.[13] Amazon initiated a one-year moratorium on police use of its facial recognition technology, calling for

---

[1] *About Face ID Advanced Technology*, Apple, Inc. (Feb. 26, 2020), https://support.apple.com/en-us/HT208108
[2] Tim Hornyak, *Smile! NEC's New Biometric Security Software Unlocks Your PC with Your Face*, PCWorld (Apr. 22, 2014), https://www.pcworld.com/article/2146660/nec-launches-facerecognition-protection-for-pcs.html.
[3] Jeff Plungis, *Car Companies Show Off Face Recognition and High-Tech Cockpit Features*, Consumer Reports (Jan. 8, 2018), www.consumerreports.org/cars-driving/car-companies-show-off-face-recognition-and-high-tech-cockpit-features/.
[4] Nick Tabor, *Smile! The Secretive Business of Facial-Recognition Software in Retail Stores*, N.Y. Mag. (Oct. 20, 2018), nymag.com/intelligencer/2018/10/retailers-are-using-facial-recognition-technology-too.html.
[5] *See, e.g., CaixaBank's ATMs with Facial Recognition, Tech Project of the Year, by The Banker,* CaixaBank (Jan. 2019), www.caixabank.com/comunicacion/noticia/caixabanks-atms-with-facial-recognition-tech-project-of-the-year-by-the-banker_en.html?id=41844.
[6] Scott McCartney, *Are You Ready for Facial Recognition at the Airport?*, Wall St. J. (Aug. 14, 2019), www.wsj.com/articles/are-you-ready-for-facial-recognition-at-the-airport-11565775008.
[7] Simson Garfinkel, *Future Tech: One Face in 6 Billion*, 23 Discover Mag. 17 (2002).
[8] Kashmir Hill, *Wrongfully Accused*, N.Y. Times (June 25, 2020), https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html; Elisha Anderson, *Controversial Detroit Facial Recognition Got Him Arrested for a Crime He Didn't Commit,* Detroit Free Press (July 10, 2020), https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/.
[9] Clare Garvie et al., Geo. L. Center for Privacy & Tech., The Perpetual Line-Up (2016).
[10] Mariko Hirose, *Privacy in Public Spaces: The Reasonable Expectation of Privacy Against the Dragnet Use of Facial Recognition Technology*, 49 Conn. L. Rev. 1591 (2017).
[11] *See, e.g., Russel Brandom, Facebook, Twitter, and Instagram Surveillance Tool Was Used to Arrest Baltimore Protestors*, The Verge (Oct. 11, 2016), https://www.theverge.com/2016/10/11/13243890/facebook-twitter-instagram-police-surveillance-geofeedia-api; *Baltimore County Police Department and Geofeedia Partner to Protect the Public During Freddie Gray Riots*, Am. C.L. Union, http://www.aclunc.org/docs/20161011_geofeedia_baltimore_case_study.pdf .
[12] *See, e.g., Facial Recognition Technology (Part 1): Its Impact on Our Civil Rights and Liberties: Hearing Before the H. Comm. on Oversight and Reform*, 116th Cong. (2019) (statement of Neema Singh Guliani, Senior Legislative Counsel, Am. C.L. Union).
[13] Arvind Krishna, *IBM CEO's Letter to Congress on Racial Justice Reform*, IBM 2 (June 8, 2020), https://www.ibm.com/blogs/policy/wp-content/uploads/2020/06/Letter-from-IBM.pdf.

"governments [to] put in place stronger regulations to govern the ethical use of facial recognition technology."[14] Microsoft, too, announced that they will not sell FRT software to police departments "until we have a national law in place, grounded in human rights."[15]

Numerous pieces of state and federal legislation in the US echo this call. Many propose a moratorium on government use of FRT until comprehensive guidelines can be set. One U.S. Senate bill proposes to bar federal agencies and federally funded programs from using FRT.[16] The state of Massachusetts has proposed restricting state usage of FRT,[17] and the City of San Francisco enacted legislation to prohibit municipal departments from using FRT.[18]

All of us support these calls for rigorous reflection about the use of FRT and one common thread throughout nearly all proposed and passed pieces of legislation is a need to understand the accuracy of facial recognition systems, within the exact context of their intended use. The federal Facial Recognition Technology Warrant Act, for example, calls for "independent tests of the performance of the system in typical operational conditions" in order to receive a warrant to use facial recognition for a given task within the government;[19] the Ethical Use of Facial Recognition Act calls for a moratorium on government use of FRT until regulatory guidelines can be established to prevent "inaccurate results";[20] the State of Washington requires that FRT vendors to enable "legitimate, independent and reasonable tests" for "accuracy and unfair performance differences across distinct subpopulations;"[21] the state of Massachusetts proposes "standards for minimum accuracy rates"[22] as a condition for FRT use in the state. The push for accuracy testing is not unique to the United States. The European Union Agency for Fundamental Rights has similarly emphasized the need to make accuracy assessments for different population groups,[23] and the European Commission emphasizes the need to demonstrate robustness and accuracy with AI systems.[24]

Understanding true in-domain accuracy — that is, accuracy of FRT deployment in a specific context — is crucial for all stakeholders to have a grounded understanding of the capabilities of the technology. FRT vendors require objective, standardized accuracy tests to meaningfully compete based on technological improvements.[25] FRT users require in-domain accuracy to acquire FRT platforms that are of highest value in the posited application. Civil society groups, academics, and the public would benefit from a common understanding of the capabilities and limitations of the technology in order to properly assess risks and benefits. Therefore, we took a concerted effort to examine this specific question

[14] *We are Implementing a One-Year Moratorium on Police Use of Rekognition*, Amazon, Inc. (June 10, 2020), https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition.
[15] Jay Greene, *Microsoft Won't Sell Police its Facial-Recognition Technology, Following Similar Moves by Amazon and IBM*, Wash. Post. (June 11, 2020), https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/.
[16] Ethical Use of Facial Recognition Act of 2020, S.B. 3284, 116th Cong. (2020).
[17] S.B. 1385, 191st Gen. Ct. (Mass. 2020).
[18] S.F., Cal., Admin. Code ch. 19B.
[19] Facial Recognition Technology Warrant Act, S.B. 2878, 116th Cong. § 5(b)(1) (2019).
[20] S.B. 3284 § 6(c)(1)(B).
[21] S.B. 6280, 66th Leg. § 6(1)(a) (Wash. 2020).
[22] Mass. S.B. 1385 § 14(b)(3).
[23] European Union Agency for Fundamental Rights, Facial Recognition Technology: Fundamental Rights Considerations in the Context of Law Enforcement (2019).
[24] European Commission, On Artificial Intelligence - A European Approach to Excellence and Trust (2020).
[25] See, e.g., Jonathan Lippman et al., Clearview AI: Accuracy Test Report (2019).

of the technology, in hopes of better understanding the operational dynamics in the field.

Although it may seem simple at first glance, understanding performance of facial recognition for a given real-world task — e.g. identifying individuals from stills of closed-circuit television video capture — is not in fact an easy undertaking. Many FRT vendors advertise stunning performance of their software.[26] And to be sure, we have witnessed dramatic advances in computer vision over the past decade, but these claims of accuracy are not necessarily indicative of how the technology will work in the field. The context in which accuracy is measured is often vastly different from the context in which FRT is applied. For instance, FRT vendors may train their images with well-lit, clear images and with proper software usage from machine learning professionals,[27] but during deployment, clients such as law enforcement may use FRT based on live video in police body cameras, later evaluated by officers with no technical training.[28] The accuracy of FRT in one domain does not translate to its uses in other domains —and changing context can significantly impact performance, as is common knowledge in the computer science literature.[29,30]

One central concern of such cross-domain performance, which has given rise to profound criticisms of FRT, is that models may exhibit sharply different performance across demographic groups. Models trained disproportionately on light-skinned individuals, for instance, may perform poorly on dark-skinned individuals.[31] A leading report, for instance, found that false positive rates varied by factors of 10 to 100 across demographic groups, with such errors being "highest in West and East African and East Asian people, and lowest in Eastern European individuals."[32]

In this White Paper, we characterize this gulf between the contexts in which facial recognition technology is created and deployed as stemming from two sources: *domain shifts* stemming from data differences across domains and *institutional shifts* in how humans incorporate FRT output in decisions. We outline concrete, actionable methods to access deployment-domain accuracy of FRT.

In our view, the ability to evaluate the accuracy of FRT is critical to the normative debates surrounding FRT. First, if a system simply does not perform as billed, and if accuracy differs dramatically across demographic groups,

[26] See, e.g., RealNetworks, SAFR Facial Recognition Platform 4 (2019), https://safr.com/wp-content/uploads/2019/08/SAFR_Platform_Whitepaper_letter_0419.2.pdf; *Face Compare SDK*, Face++, https://www.faceplusplus.com/face-compare-sdk/ (last visited June 27, 2020); Jonathan Greig, *Air Force Hires Trueface for Facial Recognition on Bases*, TechRepublic (Nov. 19, 2019), https://www.techrepublic.com/article/air-force-hires-trueface-for-facial-recognition-on-bases/; *FAQs*, Kairos, https://www.kairos.com/faq (last visited June 27, 2020); *NEC Face Recognition Technology Ranks First in NIST Accuracy Testing*, NEC Am. (Oct. 3, 2019), https://www.nec.com/en/press/201910/global_20191003_01.html.

[27] *See, e.g.*, Patrick Grother et al., Nat'l Inst. of Standards & Tech., Face Recognition Vendor Test (FRVT) Part 1: Verification 29-30 (2019) [hereinafter Grother et al., FRVT Part 1: Verification]; Gary Huang, *Labeled Faces in the Wild Home*, U. Mass. Amherst (May 2017), http://vis-www.cs.umass.edu/lfw/.

[28] Clare Garvie, Geo. L. Center for Privacy & Tech., Garbage In, Garbage Out (2019).

[29] Amos Storkey. *When Training and Test Sets are Different: Characterizing Learning Transfer*. Dataset Shift in Machine Learning 3-28 (2009) at 3-7; Vladimir Vapnik, The Nature of Statistical Learning Theory (2013); Olivier Bousquet & André Elisseef, *Stability and Generalization*. 2 J. Mach. Learn. Res. 499 (2002); H. Shimodara, *Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function*, 228 J. Stat. Planning & Inference 90 (2000); Alexey Tsymbal, Trinity Coll. Dublin, The Problem of Concept Drift: Definitions and Related Work (2004).

[30] We note that there is an entire field of AI dedicated to solving this problem, dubbed transfer learning, but to our knowledge, FRT systems do not use these techniques. *See* Sinno Jialin Pan et al., *A Survey on Transfer Learning*, 22 IEEE Transactions on Knowledge and Data Engineering 1345 (2009).

[31] Grother et al., Nat'l Inst. of Standards & Tech., Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects (2019); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. 81 Proc. Machine Learning Res. 77 (2018).

[32] Grother et al., *supra* note 31, at 2.

poor performance may disqualify an FRT system from use and obviate the need for other normative considerations. Second, performance interacts directly with normative questions. For example, lower accuracy heightens concerns about the cost of misidentification. Higher accuracy, on the other hand, amplifies concerns over surveillance, privacy, and freedom of expression. The central role of accuracy in these debates likely explains why so much proposed legislation has called for rigorous assessments of performance and is why we have tailored this White Paper to the subject.

Of course, many other considerations factor into the adoption of FRT. Concerns over privacy,[33] consent,[34] transparency,[35] and biased usage[36] all significantly complicate the use of FRT systems, independent of accuracy. While such concerns are critical to a meaningful discussion about FRT, they fall outside the direct scope of this White Paper. The scope here remains intentionally narrow, as consensus around *how* to assess the operational limits of the technology can be crafted more readily than consensus around wide-ranging normative commitments around the technology. For a broader normative assessment, each individual use case must necessarily be judged by the potential harms and benefits along all of these dimensions and we point readers to broader discussions in the references cited throughout this White Paper.

[33] No matter the application, the collection and storage of vast amounts of biometric data leads to considerable security and privacy concerns: who has access, where is the data stored, for how long, and how the data is collected all are weighted with privacy considerations. For a more thorough treatment of this issue, we refer interested readers to Gov't Accountability Office, GAO-15-621, Facial Recognition Technology: Commercial Uses, Privacy Issues, and Applicable Federal Law; *Seeing is ID'ing: Facial Recognition and Privacy*, Center for Democracy & Tech. (Jan. 22, 2012), https://cdt.org/wp-content/uploads/pdfs/Facial_Recognition_and_Privacy-Center_for_Democracy_and_Technology-January_2012.pdf.

[34] In certain contexts, meaningful consent for facial recognition is impossible to acquire: if cameras connected to FRT are ever-present in public spaces, signs warning civilians are likely to be seen too late, and the burden of finding another place to go through may be too great. Additionally, avoiding the cameras may be interpreted as an act warranting suspicion, effectively limiting any individual's ability to refuse exposure to FRT. *See, e.g.*, Lizzie Dearden, *Police Stop People for Covering Their Faces from Facial Recognition Camera Then Fine Man £90 After He Protested*, The Independent (Jan. 31, 2019), https://www.independent.co.uk/news/uk/crime/facial-recognition-cameras-technology-london-trial-met-police-face-cover-man-fined-a8756936.html. For a thorough investigation of facial recognition and consent, see Evan Selinger & Woodrow Hartzog, *The Inconsentability of Facial Surveillance*, 66 Loy. L. Rev. 101 (2019).

[35] Relatedly, public and private use of FRT to date have been handled with little transparency, as several civil society organizations, such as the EFF, ACLU, and Project on Government Oversight, were denied FOIA and RTK requests over what technologies were used for what purposes in different areas. *See Elec. Frontier Found. v. United States Dep't of Justice*, No. 4:16-cv-02041 (N.D. Cal. Apr. 19, 2016), *available at* https://www.clearinghouse.net/detail.php?id=16071; Taylor Tedford, *ICE Refuses to Turn Over Internal Documents on Facial Recognition and Tech Detention Tactics, Lawsuit Says*, Wash. Post (Nov. 7, 2019), https://www.washingtonpost.com/business/2019/11/07/ice-refuses-turn-over-internal-documents-facial-recognition-tech-detention-tactics-lawsuit-says/. Transparency and consent are tightly linked, as consent is impossible without knowledge of the systems in use.

[36] Even if FRT vendors develop systems with equal accuracies for all demographic groups, these FRT systems are nevertheless used in an imperfect world. As a result, FRT usage will likely exacerbate the already-amplified surveillance that people of color and the poor experience. We refer readers interested in biased usage to arguments presented in ACLU's letter to the House Oversight and Reform Committee. *See Coalition Letter Calling for a Federal Moratorium on Face Recognition*, Am. C.L. Union (June 3, 2019), https://www.aclu.org/letter/coalition-letter-calling-federal-moratorium-face-recognition.

2. The Challenge of Performance Assessment in New Domains

FRT vendors report stunning accuracies of their products: SAFR,[37] Kairos,[38] Face++,[39] and others report accuracies above 99% on National Institute of Standards and Technology (NIST) tests and other benchmark datasets, such as Labeled Faces in the Wild (LFW).[40] Given these reports of performance, it may seem natural to assume that FRT can take on any facial recognition challenge.

That view is wrong. Although FRT may be deployed in diverse contexts, the model is not necessarily trained to work specifically in these domains. Moreover, users may have limited understanding of model output. As a result, reported performance does not necessarily reflect model behavior and usage in a wide array of application areas.

We use *domain shift* to refer to data differences between the development and user domains.[41] On the human side, we use *institutional shift* to refer to differences in the human interpretation and usage of models across institutions, even when the data remains identical. Both domain and institutional shifts can induce large performance differences in FRT.

2.1 DOMAIN SHIFT

Domain shift arises from the difference between the types of images used by vendors and third-party auditors to train models and test performance, and the types of images used by FRT consumers to perform their desired tasks. While the datasets used by vendors are not disclosed, there is reason to believe that there are substantial differences between vendor and user images: they may have different face properties (e.g., skin color, hair color, hair styles, glasses, facial hair, age), lighting, blurriness, cropping, quality, amount of face covered, etc. Vendor and user images likely come from different distributions.

A central concept of current machine learning is that accuracy guarantees are largely domain-specific: good performance on a certain type of image data does not necessarily translate to another type of image data. This is due to the fact that machine learning models are built to recognize patterns in a certain distribution (type, set, or class) of images, and if the images fed into the model during deployment are substantially different from the images it was trained on, the patterns that the model learned may not apply, and accuracy will likely suffer.

[37] Real Networks, *supra* note 26.
[38] Kairos, *supra* note 26.
[39] Face++, *supra* note 26.
[40] Gary Huang, *Labeled Faces in the Wild Home*, U. Mass. Amherst (May 2017), http://vis-www.cs.umass.edu/lfw/.
[41] In the machine learning literature, these data differences are more precisely described as covariate shift, label shift, and concept drift. See Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla & Francisco Herrera, *A Unifying View on Dataset Shift in Classification*, 45 Pattern Recognition 521, 522-25 (2012).

Accuracy guarantees of machine learning models depend upon the similarity of training, testing and deployment images.[42] If images fall outside of the training/test data distribution, we have little sense for how well the model will perform.

While FRT distributors do not disclose the makeup of their training sets publicly, there is some evidence to suggest that the most common practice is to train models based on images scraped from individuals' photos on the internet published with a creative commons license, such as on Flickr.[43] These datasets are commonly regarded as not similar to several real-life deployment domains. Subjects are, for the most part, aware that their pictures are being taken intentionally, and as a result, these pictures are often clear and well-lit. While some datasets have been created to mimic contexts with unsuspecting photo subjects and low-lighting domains,[44] the size of these datasets is much smaller than those scraped off the internet. Moreover, subject skin color, hair style, age, etc. may not reflect those in a new application. Without domain-specific training data, it is unlikely that the accuracy reported by FRT vendors applies to in-domain use.

In addition to the likely differences between training imagery and in-domain imagery, it is known that the images on which FRT are *benchmarked* are distinct. The datasets used by FRT vendors to report accuracies, such as Labeled Faces in the Wild (LFW), consist of images that are vastly different than those coming from most of the application domains where FRT is actually used. The LFW creators themselves acknowledge these limitations, noting that "no matter what the performance of an algorithm on LFW, it should not be used to conclude that an algorithm is suitable for any commercial purpose," as the dataset has "a relatively small proportion of women . . . [and] many ethnicities have very minor representation or none at all," and that additionally, "poor lighting, extreme pose, strong occlusions, low resolution . . . do not constitute a major part of LFW" which thus disqualifies the dataset for use as a commercial benchmark.[45] While NIST has done leading work to benchmark FRT systems, NIST benchmark datasets are still substantially more controlled than many applications. The NIST dataset of "in the wild" imagery consists of "unconstrained photojournalism and amateur photography imagery" that were "cropped prior to passing them to the algorithm" to be tested.[46] The rest of the images considered in that study were mugshot images of varying quality, which are also often centered on the face of a cooperative subject.[47]

In short, there are strong reasons to believe that domain shift creates the potential for serious performance degradation in new domains.

---

[42] Vapnik, *supra* note 29; Bousquet & Elisseef, supra note 29. As a caveat, we note that there are some ways to adapt an FRT model to perform well on a set of related, but different images than the group it was trained on: this relies on methods from the discipline of domain adaptation. As far as the authors are aware, FRT vendors often do not take advantage of *domain adaptation* techniques to fine-tune their models to consumer use-cases. However, we note that even if vendors did utilize domain adaptation, there still is no substitute for training a model on the correct deployment domain.

43 Madhumita Murgia, *Who's Using Your Face? The Ugly Truth About Facial Recognition*, Fin. Times (Sept. 18, 2019), https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e.

44 *See, e.g.*, Adam Harvey & Jules LaPlace, *MegaPixels: Origins, Ethics, and Privacy Implications of Publicly Available Face Recognition Image Datasets*, MegaPixels (2020), https://megapixels.cc/.

45 Huang, *supra* note 27.

46 Patrick Grother et al., Nat'l Inst. of Standards & Tech., Face Recognition Vendor Test (FRVT) Part 2: Identification 16 (2019) [hereinafter Grother et al., FRVT Part 2: Identification].

47 *Id*. at 14-15.

2.2 INSTITUTIONAL SHIFT

Performance differences may also arise from institutional shifts in deployment. The understanding of technological tools, such as FRT, may be "inseparable from the specifically situated practices of their use."[48] As articulated by Green and Chen,[49] the performance of AI systems is often understood through statistical metrics of accuracy, but technical accuracy does not reflect the true effect that the technology has in the field because humans still typically act on that technology.

Such performance differences can hence arise even with identical imagery by vendors and users. FRT could cause two institutions deploying identical systems on identical imagery (e.g., two police departments in adjacent jurisdictions) to diverge and exhibit sharply different operational performance. One specific example of this lies in the use of confidence scores. While Amazon Rekognition recommends a 99% confidence threshold on identity matching for use in law enforcement applications,[50] one sheriff's office reported, for instance, "We do not set nor do we utilize a confidence threshold."[51] Operational performance would likely be quite different for a department that abides by Amazon's recommendations.

Much research documents the potential divergence between raw model output and human decisions based on that output. Joint human-AI system breakdown can stem from several reasons. Users may ignore model output, either because they do not understand or trust the system, or view themselves as more qualified, as some experiments with judges using pretrial risk assessment algorithms suggest.[52] Alternatively, users may over-trust the algorithm, as documented by experiments finding that users over-trust a system billed as accurate, even if it clearly gives no useful information (e.g. generates random outputs).[53] Users have also been shown to selectively listen to model output that confirms their own biases, which can lead to amplified discrimination concerns.[54]

Where FRT is embedded in a human system, understanding its performance hence also requires understanding the impact on human decision makers. Such performance measurement would be particularly valuable to incentivize best practices (e.g., training, communication) in the adoption of FRT, rather than mandating specific practices.[55]

---

[48] Lucy Suchman et al., *Reconstructing Technologies as Social Practice*, 43 Am. Behav. Scientist 392, 399 (1999).
[49] Ben Green & Yiling Chen, *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments*, 2019 Proc. of the Conf. on Fairness, Accountability, and Transparency 90, 90 [hereinafter Green & Chen, *Disparate Interactions*].
[50] *Use Cases that Involve Public Safety*, Amazon, Inc. (2020), https://docs.aws.amazon.com/rekognition/latest/dg/considerations-public-safety-use-cases.html.
[51] Bryan Menegus, Defense of Amazon's Face Recognition Tool Undermined by Its Only Known Police Client, Gizmodo (Jan. 31, 2019), https://gizmodo.com/defense-of-amazons-face-recognition-tool-undermined-by-1832238149. *See also* Jennifer Lynch, Elec. Frontier Found., Face Off: Law Enforcement Use of Face Recognition Technology 15-16 (2020).
[52] *See, e.g.*, Megan Stevenson, *Assessing Risk Assessment in Action*, 103 Minn. L. R. 303 (2018); Megan T. Stevenson & Jennifer L. Doleac, The Am. Const. Soc'y, The Roadblock to Reform (2018); Angèle Christin, *Algorithms in Practice: Comparing Web Journalism and Criminal Justice*, 4 Big Data & Soc'y 1 (2017).
[53] *See, e.g.*, Birte Englich et al., *Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making*, 32 Personality and Soc. Psychol. Bull. 188 (2006); Aaron Springer et al., *Dice in the Black Box: User Experiences with an Inscrutable Algorithm*, Cornell U. (Dec. 7, 2018), https://arxiv.org/abs/1812.03219.
[54] *See, e.g.*, Green & Chen, *Disparate Interactions, supra* note 45; Bo Cowgill, The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities (Dec. 5, 2018) (unpublished manuscript) (on file with author); Alex Albright, If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions (Sept. 3, 2019) (unpublished manuscript) (on file with author); Jennifer Skeem et al., *Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants*, 51 L. & Hum. Behav. (2020).
[55] For a classic articulation of the benefits of performance-based regulation over command-and-control systems, see Bruce A. Ackerman & Richard B. Stewart, *Reforming Environmental Law*, 37 Stan. L. Rev. 1333 (1985).

3. Recommendations for a Protocol

Given the challenges of domain and institutional shifts, we provide recommendations to facilitate more rigorous evaluation of in-domain (operational) performance.

Our recommendations are grounded by three principles. First, we build on what is known in the technical literature about domain shift and domain adaptation, as well as in the interdisciplinary work on human-computer interaction. That said, we recognize that this research is rapidly advancing, so we refrain from advocating any specific technical solution that may soon be superseded. Instead, our goal is to provide a general protocol that can enable more rigorous and widespread assessment of in-domain performance independent of specific technical details. Second, we develop recommendations that are meaningful in advancing an understanding of in-domain performance, but can also plausibly be implemented in the near term.[56] Third, our recommendations encompass all potential stakeholders, with the goal of empowering not only vendors and users, but also the public sector, third-party auditors, academics, and civil society to participate in developing a more grounded understanding of FRT in operation.

The first set of recommendations address the data-centered roadblocks to establishing reliable in-domain accuracy, and the second set of recommendations focus on evaluating the human component affecting in-domain FRT system accuracy.

3.1 DATA-CENTERED RECOMMENDATIONS

A major roadblock to rigorous assessment of in-domain FRT performance is that too little is known about the imagery on which a model was built. And even if users wanted to assess domain shift, not all systems readily enable such testing. We hence provide a protocol that would enable such assessments.

3.1.1 Vendor and Third-Party Data Transparency

Comprehensive Publication of Imagery. Vendors and third parties should be transparent about training and testing imagery. When using public datasets, vendors and third parties should maintain an up-to-date list of the datasets used for each software release. For private datasets, parties should disclose the full training/testing data, along with documentation,[57] which enables users to compare images to assess the potential for domain shift.

Fallback of Random Sample and Comparison Metrics. Although publishing the full imagery is the ideal solution,[58] a less desirable fallback would be the publication of a large random sample of imagery[59] and enabling the comprehensive calculation within the

[56] For a more ambitious proposal that proposes a new federal regulatory agency specifically for FRT, see Erik Learned-Miller, Vicente Ordonez, Jamie Morgenstern & Joy Buolamwini, Facial Recognition Technologies in the Wild: A Call for a Federal Office (2020).
[57] *See* Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III & Kate Crawford, *Datasheets for Datasets* (Mar. 23, 2018), *arXiv preprint arXiv:1803.09010*.
[58] Releasing data may be infeasible due to other constraints (e.g., privacy, intellectual property).
[59] To prevent the gaming of disclosed imagery, users might be able to select the random images themselves. First, the vendor or third party could provide a public hash of each of their images. The hash reveals nothing about the images and does not give away any company secrets. Second, the user, when evaluating a system, would supply a list of random image hashes. Third, the vendor or third party would reveal those images from their training data. This process will have to be rate-limited to prevent scraping of the underlying dataset.

vendor's system of comparison metrics between training and user data (or an FRT model's internal representation of the training data and the user's data).[60] It is possible the calculation of such metrics would enhance the user's ability to assess image differences, but further research is required to understand the utility of this approach.

3.1.2 Facilitating In-Domain Testing

Although comparing dataset distributions will provide information about how distant training and deployment domains are, it does not provide a rigorous assessment of in-domain performance.

Enabling Testing. Vendors and users should hence facilitate independent validation of in-domain performance, with programmatic access within their platform. The principal method to assess in-domain accuracy lies in *labeled data* from the application domain when available — i.e. image inputs with ground truth labels (e.g., identity of individual) for the user's specific application. For example, to test for accuracy in identifying individuals who should have building access, one would apply the model to images from the same building camera and compare model predictions with ground truth labels of individuals with building access. This procedure should be enabled within vendor systems.

Labeling Interface. A common problem, however, is the lack of in-domain labeled datasets. To facilitate the creation of such labeled datasets, we recommend that vendors enable users to label their own data and test the vendor's facial recognition models using that data. This can be either a labeling interface (e.g., Amazon Rekognition Custom Labels), or the ability for users to upload pre-labelled data. In addition, vendors should provide programmatic access (e.g., via API) to enable users to assess performance with user-provided, domain imagery. To provide independent assessments, users should ideally reserve holdout testing data and define acceptable metric thresholds that must be met prior to operational deployment.

3.1.3 Performance Assessment Over Time

A one-time accuracy check may still be insufficient. Domain shift can also enter the system via changes to the dataset distribution over time[61] or vendor software updates.

Documentation. Vendors should hence provide detailed release notes and documentation for each version of the FRT system, including changes in the model and data. We recommend the changes to be described in as much detail as possible, although we recognize that exact model architectures may be deemed proprietary knowledge. Release notes should also include changes to training data, training algorithms, parameters, fairness constraints and any other aspects that might influence performance. Any such changes should be considered a new release, with its own release notes, which in turn might trigger user recertification.[62]

---

[60] Commonly used measures for domain discrepancy include measures like Maximum Mean Discrepancy (MMD), Arthur Gretton et al., *A Kernel Two-Sample Test*, 13 J. Mach. Learn. Res. 723 (2012), Kullback-Leibler Divergence, Wouter M. Kouw et al., *An Introduction to Domain Adaptation and Transfer Learning*, Cornell U. (Dec. 31, 2018), https://arxiv.org/abs/1812.11806, and Hellinger Distance, Gregory Ditzler & Robi Polikar, *Hellinger Distance Based Drift Detection for Nonstationary Environments*, 2011 IEEE Symp. on Computational Intelligence in Dynamic and Uncertain Environments. There are several challenges to address in applying such metrics: deciding upon what aspects of the data the metric focuses on (e.g. KL divergence can be calculated over image brightness, color, etc., or a combination of many aspects); whether the metric is evaluated on raw or transformed data, and how to enforce uniformity in the application of the metric, given these considerations.

[61] Alexey Tsymbal, Trinity Coll. Dublin, The Problem of Concept Drift: Definitions and Related Work (2004).

[62] We note that these recommendations are similar in spirit to the certification ideas in Erik Learned-Miller, Vicente Ordonez, Jamie Morgenstern & Joy Buolamwini, Facial Recognition Technologies in the Wild: A Call for a Federal Office (2020).

Such documentation about model and data changes over time would facilitate an assessment of domain drift over time, such as recertification with any data and model updates.

3.2 HUMAN-CENTERED RECOMMENDATIONS

While evaluating in-domain accuracy is necessary to understanding operational performance, technical accuracy alone remains inadequate. FRT outputs are used by humans, deployed within existing institutions, and interact in a social setting. Understanding the human-FRT interaction is hence an integral part of evaluating in-domain FRT performance. "[S]tatistical properties (e.g., AUC and fairness) do not fully determine [an AI tool's] impacts when introduced in social contexts."[63] Even though most computer vision research focuses on optimizing accuracy isolated from its human surroundings, we urge an assessment that encompasses accuracy in context.[64]

The most direct approach to test in-domain operational accuracy of FRT lies in a field experiment of actual usage. For example, in a criminal justice context, this would involve assessing the human accuracy of identifications in instances when the FRT system is used versus when it is not. A prominent example lies in a field experiment in predictive policing, wherein the output of predictive crime models were randomly disclosed to police districts and outcomes were compared across districts.[65] (The trial found no statistically significant evidence of crime reduction from predictive policing.) As is widely recognized in the social sciences, field experiments overcome external validity concerns of laboratory experiments and thorny causal identification challenges of observational studies.[66] That said, such end-to-end field experiments can be resource-intensive to design and administer[67] and pose potential ethical challenges and risks to a community.[68] Moreover, in the FRT context, there are significant questions about what objective outcome measures are available. In the police identification context, for instance, arrests and convictions may themselves be affected by bias and/or perceptions of FRT. If convictions increased, that may not be reflective of FRT performance, as much as human overreliance on automated systems.

While efforts to conduct field experiments of technological adoption are laudable, in order to facilitate more widespread testing, we recommend testing of the impact of FRT on more immediately observable (i.e., surrogate) outcomes. Such tests may not allow one to infer the effects of FRT on ultimate outcomes (e.g., crime rates), but they enable the assessment of key mechanisms by which technology may affect human decision making.

To illustrate, consider well-known breakdowns in human-machine interaction. Some users may *over-rely* on machine output, sometimes dubbed "automation bias."[69]

---

[63] Ben Green & Yiling Chen, *The Principles and Limits of Algorithm-in-the-Loop Decision Making*, 3 Proc. ACM Hum.- Comput. Interact. 1, 2 (2019) [hereinafter Green & Chen, *Principles and Limits*].
[64] Bryan Wilder et al., *Learning to Complement Humans*, Cornell U. 1 (May 1, 2020), https://arxiv.org/abs/2005.00582.
[65] Priscillia Hunt, Jessica Saunders & John S. Hollywood, Evaluation of the Shreveport Predictive Policing Experiment (2014). See also an in-domain assessment of fire prediction software: Jessica Lee et al. *A Longitudinal Evaluation of A Deployed Predictive Model of Fire Risk*. AI for Social Good Workshop at the 32nd Conf. Neural Info. Processing Sys. (2018).
[66] Alan S. Gerber & Donald Green, Field Experiments: Design, Analysis, and Interpretation (2012).
[67] *But see* Cassandra Handan-Nader, Daniel E. Ho & Becky Elias, *Feasible Policy Evaluation by Design: A Randomized Synthetic Stepped-Wedge Trial in King County*, 44 Evaluation Rev. 3 (2020).
[68] See Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals (Scott Desposato ed. 2015).
[69] Linda J. Skitka et al., *Does Automation Bias Decision-Making?*, 51 Int'l J. Hum.-Computer Studies 991 (1999); Jeffrey Warshaw et al., *Can an Algorithm Know the "Real You"? Understanding People's Reactions to Hyper-personal Analytics Systems*, Proc. of the 33d Ann. ACM Conference on Hum. Factors in Computing Sys. 797 (2015); *see also* Englich et al., *supra* note 49; Springer et al., *supra* note 49.

In certain enforcement contexts, FRT face matches are trusted without regard to reported system accuracy or confidence of output.[70] Over-trusting machine outputs can lead to a drop in operational performance, as suboptimal predictions from the algorithm are acted upon. On the other hand, some may *under-rely* on machine output, sometimes dubbed "algorithm aversion."[71] In the criminal justice context, for instance, some judges entirely ignore risk assessment scores.[72] And yet others may *selectively rely* on machine outputs depending on prior beliefs and biases. Some evidence suggests that judges, for instance, give harsher sentences to black defendants than white defendants with moderate risk scores.[73]

We hence recommend, wherever possible, pilot A/B testing to assess the impacts of the FRT output on specific human decisions. When an FRT system delivers model output to human decisionmakers, A/B testing would randomize elements of FRT output to assess the effect on human decisions. (In a sense, these A/B tests are still "field experiments," but in contrast to ambitious designs that attempt to assess the impact of FRT on crime, these tests focus on immediately observable surrogate outcomes.) We offer several examples in the context of police lineups, where candidate images, potentially selected by FRT, are presented for human identification.

- An A/B test could randomly withhold (or randomly disclose) whether the candidate imagery was selected by an FRT system. If the disclosure causes greater willingness to infer a match, that provides evidence of over-reliance on FRT. Conversely, if the disclosure causes lower willingness to infer a match, that provides evidence of algorithm aversion.

- Another A/B test could randomly shuffle the confidence scores of FRT results to assess whether users appropriately incorporate uncertainty into their decisions. Such a design would enable the assessment of *responsiveness* to confidence scores, as well as selective responsiveness along demographic attributes (e.g., race and gender). It may also be indicative of improper training of users and/or improper calibration of confidence scores by vendors.

- Another A/B test would reserve a random holdout of decisions to be determined based on the pre-existing (non-FRT) system. A comparison between the manual and FRT-augmented decisions would enable an assessment of the effects of the FRT system on performance.[74]

We recognize that these examples only scratch the surface of the full impact of FRT on human decisions across all contexts. Not all outcomes can be studied with this approach. The impact of mediating variables that exist only at the institution-level (e.g., managerial oversight, budget) cannot easily be assessed without many A/B tests across institutions. But the more A/B pilot tests become standard practice as FRT systems are adopted, the more we will be able to ground operational performance and accuracy. Much work remains to be done to understand the effects of FRT's deployment within institutions.

---

[70] *See* Garvie, *supra* note 28; Bryan Menegus, *Amazon's Defense of Rekognition Tool Undermined by Police Client*, Gizmodo (Jan. 31, 2019), https://gizmodo.com/defense-of-amazons-face-recognition-tool-undermined-by-1832238149.
[71] Berkeley J. Dietvorst et al., Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err, 144 J. Experimental Psychol.: General 114 (2015).
[72] Brandon L. Garrett & John Monahan, *Judging Risk*, 108 Cal. L. Rev. 439 (2020).
[73] Albright, *supra* note 50, at 4.
[74] See David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 Yale J. on Reg. 800 (2020).

4. Responsibilities Beyond the Protocols

We now spell out several other recommendations for implementation of in-domain testing. Users and vendors have particular responsibilities in adopting and implementing the protocols. Opening up the FRT ecosystem to facilitate in-domain accuracy testing, however, will also empower a much wider range of parties and stakeholders to rigorously assess the technology. By opening up information about private training and testing sets and enabling in-domain testing via programmatic access, researchers, auditors, and other groups will be empowered to test suitability for different domains.

4.1 SERVICE MODEL

The recommendations of this White Paper may ultimately not be achievable through sale of FRT systems as "off-the-shelf" technology. Instead, comprehensive assessment of in-domain performance may require a shift of the business model toward an on-going service, whereby vendors collaborate on an ongoing basis with users to ensure that the system performs as desired. Such a shift toward a service model may also facilitate an improved understanding of imagery differences, model changes, constraints of use cases, and training of users.

4.2 USERS AND PROCUREMENT

A compelling lever for implementing the above protocol lies in the procurement process. When businesses and government agencies procure FRT systems through large-scale contracts, such procurement should be conditioned on rigorous in-domain accuracy tests. Users should not rely solely on NIST benchmarks that may not reflect performance in the domain for which an FRT system is procured. Instead, users should insist on compliance with the protocol spelled out in this White Paper and demand evidence for performance in the user's specific domain.

The procurement process can, of course, be complicated, and some development may be required prior to being able to test in-domain accuracy. In those settings, an intermediate solution may lie in sequencing the procurement process to conduct pilot studies to assess in-domain accuracy. U.S. Customs and Border Protection (CBP), for instance, compared the use of FRT, iris-scanning, and fingerprinting technologies for identification during border crossings.[75] Similarly, a pilot could compare different vendors, as the New York Police Department did during a trial period with predictive policing.[76] The CBP example also illustrates that a pilot may seek to answer the broader question of whether FRT is appropriate in light of available alternatives.

[75] U.S. Customs and Border Protection, Southern Border Protection Pedestrian Field Test: Summary Report (2016), https://epic.org/foia/dhs/cbp/biometric-entry-exit/Southern-Border-Pedestrian-Field-Test-Report.pdf.
[76] N.Y. Police Dep't, *Predictive Policing Pilot Evaluation*, Brennan Center for Just. (June 2016), https://www.brennancenter.org/sites/default/files/Predictive%20Policing%20Final%20802-818%20-%20%28%23%20Legal%208799488%29.PDF. Note this document was obtained through a FOIA request from the Brennan Center for Justice.

4.3 AUDITORS

We recommend that auditors expand their testing datasets to cover high-priority emerging domains.

Independent audits serve an important role in validating FRT systems. The best-known benchmarking standard is provided by NIST and its Facial Recognition Vendor Test (FRVT). Over the past twenty years, FRVT has benchmarked algorithms for performance in facial identification, facial verification, and other tasks.[77] Vendors submit an executable version of their algorithm to NIST, which NIST deploys on its (non-public) datasets to compute the algorithm's accuracy. While NIST's benchmarking is exemplary, NIST's datasets do not yet represent the wide array of application domains in which FRT systems are used. Moreover, such audits should be conducted each time the data and model are updated.

Third-party (non-government) certifications can also play an important role, but it will be important to design such processes to be independent and free from conflicts of interest.[78]

4.4 ACADEMICS

While academics have played a critical role in surfacing errors and biases in a range of AI systems — including recidivism prediction,[79] predictive policing,[80] medical diagnostics,[81] and FRT systems[82] — and spearheading basic FRT research,[83] more research remains to be done on domain and institutional shift in FRT. The current closed ecosystem has likely prevented rigorous academic scrutiny and the above protocols should enable academic researchers to engage in more rigorous assessments — as third-party evaluators in collaboration with FRT users — of the performance across unchartered domains.

4.5 MEDIA AND CIVIL SOCIETY

Media and civil society organizations have similarly had major effects on the discussion around the use of AI in public-facing contexts[84] and FRT in particular.[85] With expanded access to vendor and user information, investigative journalists and public interest groups may amplify their ability to ground our understanding of FRT performance.

[77] *See, e.g.*, Grother et al., FRVT Part 1: Verification, *supra* note 27; Grother et al., FRVT Part 2: Identification, *supra* note 42; Grother et al., Nat'l Inst. of Standards & Tech., Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects (2019).
[78] See Jodi L. Short & Michael W. Toffel, *The Integrity of Private Third-Party Compliance Monitoring*, 42 Admin. L. & Reg'y N. 22 (2016).
[79] *See*, e.g., Green & Chen, *Disparate Interactions*, *supra* note 45; Cowgill, *supra* note 50; Albright, *supra* note 50.
[80] Kristian Lum & William Isaac, *To Predict and Serve?*, 13 Significance 14 (2016).
[81] Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 Sci. 447 (2019).
[82] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. 81 Proc. Machine Learning Res. 77 (2018).
[83] *See, e.g.*, Yutian Lin et. al., *A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification*, 33 Proc. of the AAAI Conf. on Artificial Intelligence 8738 (2019).
[84] *See, e.g.*, Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[85] *See, e.g.*, Jacob Snow, *Amazons Face Recognition Falsely Matched 28 Members of Congress with Mugshots*, Am. C.L. Union (Aug. 12, 2019), www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

5. Conclusion

The debate surrounding FRT raises profound questions about technology in society. Our goals here have been limited. We have not answered the broader ethical and normative questions about FRT's impact on privacy, speech, and the power of the state. Instead, we have sought to make concrete a general requirement that appears in nearly every proposed legislation to regulate FRT: whether it works as billed.

We believe that adopting the recommendations above — by regulation, contract, or norm — will do much to improve our understanding, reflection, and consideration of one of the most important emerging applications of AI today.

Summary of Recommendations

Our recommendations focus on two sources of performance differences between the development and deployment contexts. The first source lies in *data differences* between the development and deployment domains. For instance, FRT models trained on driver's license pictures may not translate into good performance with pictures containing a wider range of positions and lighting.

RECOMMENDATION (1): PROTOCOL FOR DOMAIN-SPECIFIC MODEL ASSESSMENT

Transparency of Imagery. Vendors should be transparent about the domain of training data at all points.

> The ideal disclosure would consist of the full vendor training and test set imagery. Such transparency enables users to compare vendor training images with images in a new domain.
>
> If the full imagery set cannot be disclosed, a less desirable alternative is that vendors could disclose large random samples of imagery and facilitate the calculation of comparison metrics that summarize domain discrepancy between vendor images and user images.

Enabling Testing. Vendors and users should facilitate and conduct independent validation of in-domain performance.

> First, vendors should provide programmatic access (e.g., via API) to enable users and third parties to assess performance with new domain imagery. Such access should ideally also enable users to label their own testing data, which is required for in-domain performance assessment.
>
> Second, users can reserve holdout testing data and define acceptable metric thresholds that must be met prior to operational deployment.
>
> Third, to protect against temporal changes and to ensure that changes in the vendor's model do not adversely affect performance, vendors and users should enable and conduct periodic recertification of performance.

Documentation. Vendors should provide comprehensive release notes and documentation for each model version.

> The release notes should, at minimum, include any changes to underlying model and architecture; performance metrics across subcategories such as demographics and image quality; and information about training or evaluation data. Such documentation would facilitate an assessment of temporal changes and potentially trigger recertification.

The second source for performance differences between development and deployment stems from *institutional differences* that heighten the discrepancy between vendor-reported and operational accuracy. Diverse institutional contexts can induce common problems in human-computer interaction: users, for instance, may over-rely on model output (e.g., adhering to FRT output even when clearly wrong) or selectively use model output in a way that exacerbates demographic bias (e.g., overriding system suggestions for one race, but not another).

RECOMMENDATION (2): PROTOCOL FOR EVALUATING THE IMPACT OF FRT ON HUMAN DECISIONS

Users should test the specific effects of FRT on elements of human decision making where possible.

> While a rigorous evaluation of the impact of the FRT system on human decision making can be complex, pilot A/B tests that are conventional in web platforms can be adapted to assess the specific effects of FRT output on human decisions. For instance, withholding confidence scores (which indicated the confidence of identification) for a random subset of images may enable an assessment of overreliance and the potential for selective reliance.

Our last set of recommendations focuses on implementation, procurement, and auditing.

RECOMMENDATION (3): PROCUREMENT AND AUDITING

Procurement. Users, including governments and companies, should condition the procurement of FRT systems on in-domain testing and adherence to the protocol articulated above. To provide a comprehensive sense of in-domain performance, the procurement process should include an assessment of (a) technical accuracy and (b) the effects of FRT on human decision-making.

Expanding Auditing. With the protocols spelled out above implemented, third parties, academics, journalists, and civil society organizations should consider expanding performance benchmarks to audit systems on a wider range of domain imagery.

Concurring and Dissenting View

Note: In the course of circulating the draft White Paper, we have taken note of a 'dissenting' perspective pertaining to the scope of the Paper. In full candor, we attempt to provide that perspective here, which focuses on the harm of FRT.

In January 2020, police pulled up to the driveway of Robert Williams's Detroit home and handcuffed him in front of his wife and two children. He was held overnight and interrogated, missing work for the first time in four years. Detectives showed him surveillance footage from a shoplifting camera. "This is not me," Williams responded. "You think all Black men look alike?" According to the *New York Times*, Williams "may be the first known account of an American being wrongfully arrested based on a flawed match from a facial recognition algorithm."[86] The harm has been long-lasting. Since the "humiliating" experience, the Williamses have considered sending their young daughters to therapy.

The human costs of technology, borne disproportionately by vulnerable communities, must be at the foreground when we consider FRT. The NIST 99% accuracy statistic alone does not convey the gaping accuracy disparities across demographic groups, causing disproportionate harm to women, minorities, and those unable to protect themselves from intrusive expansions of surveillance.

By focusing on operational performance, the White Paper misses the broader context of how FRT is used to harm. The use by law enforcement of FRT should be categorically banned.

Questions of operational performance are a side show to deeper questions about when we ban technology that harms.

[86] Hill, *supra* note 8.

Appendix: Workshop Attendees

We are grateful to the generosity of attendees who attended the HAI workshop "Facial Recognition Technology, Measurement & Regulation Workshop" in May 2020. Affiliations are provided for identification purposes only and views expressed in the White Paper are those of the authors alone.

Hartwig Adam (Google)
Elizabeth Adams (Stanford University)
Maneesh Agrawala (Stanford University)
Eric Badiqué (European Commission)
Wendy Belluomini (IBM)
Joy Buolamwini (MIT)
Stephen Caines (Stanford University)
Rama Challeppa (University of Maryland)
Lisa Dyer (Partnership on AI)
John Etchemendy (Stanford University)
John Paul Farmer (City of New York)
Li Fei-Fei (Stanford University)
Clare Garvie (Georgetown University)
Patrick Grother (NIST)
Nick Haber (Stanford University)
Kristine Hamann (Prosecutors' Center for Excellence)
Tatsunori Hashimoto (Stanford University)
Jonathan Hedley (Amazon)
Daniel E. Ho (Stanford University)
Richard Hung (U.S. GAO)
Benji Hutchinson (NEC America)
Jen King (Stanford University)
Erik Learned-Miller (University of Massachusetts at Amherst)
Mark Lemley (Stanford University)
Iacopo Masi (University of Southern California)
Chintan Mehta (Wells Fargo)
Joshua New (IBM)
Juan Carlos Niebles (Stanford University)
Pietro Perona (Amazon)
Rob Reich (Stanford University)
Florian Schroff (Google)
Michael Sellitto (Stanford University)
Amarjot Singh (Stanford University)
Jacob Snow (American Civil Liberties Union)
Elham Tabassi (NIST)
Kate Weber (Google)
Harry Wechsler (George Mason University)
Wojciech Wiewiórowski (European Commission)
Stefanos Zafeiriou (Imperial College London

HAI Rapporteurs and Participants

Nazila Alasti (Stanford University)
Deep Ganguli (Stanford University)
Danielle Jablanski (Stanford University)
Marisa Lowe (Senate Foreign Relations Committee)
Evani Radiya-Dixit (Stanford University)
Russell Wald (Stanford University)
Christopher Wan (Stanford University)