

EVALUATING FACIAL RECOGNITION TECHNOLOGY: A PROTOCOL FOR PERFORMANCE ASSESSMENT IN NEW DOMAINS

DANIEL E. HO, EMILY BLACK, MANEESH AGRAWALA, & LI FEI-FEI[†]

ABSTRACT

Facial recognition technology (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 merit serious debate. Yet one requirement common to proposed legislation and regulation of FRT is the testing and assessment of operational performance: how well does FRT actually work? This poses deep challenges given the rapid uptake of FRT in many new domains, such as retail, finance, travel, and criminal justice. In this Article, we provide research- and science-grounded recommendations for how to concretely test the operational accuracy of FRT that will be central to regulation and oversight.

TABLE OF CONTENTS

INTRODUCTION	754
I. THE CHALLENGE OF PERFORMANCE ASSESSMENT IN NEW DOMAINS	759
<i>A. Domain Shift</i>	760
<i>B. Institutional Shift</i>	762
II. RECOMMENDATIONS FOR A PROTOCOL	763
<i>A. Data-Centered Recommendations</i>	764
1. Vendor and Third-Party Data Transparency	764
a. Comprehensive Publication of Imagery.....	764

[†] Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law at Stanford Law School, Professor of Political Science, Senior Fellow at the Stanford Institute for Economic Policy Research, Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence, and Director of the Regulation, Evaluation, and Governance Lab (RegLab) at Stanford University. Emily Black is a Ph.D. student in computer science at Carnegie Mellon University. Maneesh Agrawala is the Forest Baskett Professor of Computer Science and Director of the Brown Institute for Media Innovation at Stanford University. Li Fei-Fei is the Sequoia Professor in the Computer Science Department at Stanford University and the Denning Co-Director of Stanford’s Human-Centered AI Institute. We thank Christopher Wan, Evani Radiya-Dixit, Ohad Fried, Danielle Jablonski, and Marisa Lowe for terrific research assistance with this Article. This Article was previously released online as a White Paper for Stanford’s Institute for Human-Centered Artificial Intelligence. Co-signatories included Elizabeth Adams, Stephen Caines, John Etchemendy, Iacopo Masi, Erik Learned-Miller, and Harry Wechsler. We are also grateful for the generosity of workshop participants at the Workshop on Facial Recognition Technology, Measurement, and Regulation, Celia Clark, Russell Wald, Michael Sellitto and Nazila Alasti for organizing the workshop, and helpful comments from participants at the Georgetown Technology Law and Policy Colloquium and Joy Buolamwini, Julie Cohen, Amanda Coston, John Etchemendy, Tatsu Hashimoto, Peter Henderson, Erik Learned-Miller, Michael Sellitto, and Christine Tsang.

b. Fallback of Random Sample and Comparison Metrics.....	764
2. Facilitating In-Domain Testing	765
a. Enabling Testing	765
b. Labeling Interface	765
3. Performance Assessment Over Time	766
a. Documentation	766
<i>B. Human-Centered Recommendations</i>	766
III. RESPONSIBILITIES BEYOND THE PROTOCOL	769
<i>A. Service Model</i>	769
<i>B. Users and Procurement</i>	769
<i>C. Auditors</i>	770
<i>D. Academics</i>	771
<i>E. Media and Civil Society</i>	771
CONCLUSION	771
APPENDIX: SUMMARY OF RECOMMENDATIONS	772

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 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 airport⁷ or being falsely accused of a crime, as was the case for Robert Williams and Michael Oliver.⁸

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, 5:54 AM), <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 REPS. (Jan. 8, 2018), <https://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), <https://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. 8, 2019), https://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, 8:58 AM), <https://www.wsj.com/articles/are-you-ready-for-facial-recognition-at-the-airport-11565775008>.

7. Simson Garfinkel, *One Face in 6 Billion*, 23 DISCOVER MAG. 17, 17 (2002).

8. Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. TIMES (Aug. 3, 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 11, 2020, 11:03 PM), <https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/>.

A 2016 study reports that one out of two Americans are involved in a “perpetual line-up” (i.e., an ongoing virtual police line-up), 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 our right to assemble.¹¹

The widespread use of FRT in high-stakes contexts has led to loud calls 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 or analysis software,” stating that “now is the time to begin a national dialogue on whether and how [FRT] 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 FRT, calling for “governments [to] put in place stronger regulations to govern the ethical use of [FRT].”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 United States echo these calls.¹⁶ Many propose a moratorium on government use of FRT until comprehensive guidelines can be set.¹⁷ One U.S. Senate bill proposes to bar federal agencies and federally funded programs from using FRT.¹⁸ The State of Massachusetts has proposed restricting state usage of FRT,¹⁹

9. CLARE GARVIE, ALVARO M. BEDOYA, & JONATHAN FRANKLE, GEO. L. CTR. ON PRIV. & TECH., *THE PERPETUAL LINE-UP* 8 (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, 1611–15 (2017).

11. See, e.g., Russel Brandom, *Facebook, Twitter, and Instagram Surveillance Tool Was Used to Arrest Baltimore Protestors*, THE VERGE (Oct. 11, 2016, 1:42 PM), <https://www.theverge.com/2016/10/11/13243890/facebook-twitter-instagram-police-surveillance-geofeedia-api>; GEOFEEDIA, *BALTIMORE COUNTY POLICE DEPARTMENT AND GEOFEEDIA PARTNER TO PROTECT THE PUBLIC DURING FREDDIE GRAY RIOTS* (Oct. 11, 2016).

12. See, e.g., *Facial Recognition Technology: Part I Its Impact on Our Civil Rights and Liberties: Hearing Before the H. Comm. on Oversight & Reform*, 116th Cong. 9 (2019) (statement of Neema Singh Guliani, Senior Legislative Counsel, Am. C.L. Union).

13. Letter from Arvind Krishna, IBM Chief Exec. Officer, to Senators Kamala Harris & Cory Booker, and Representatives Karen Bass, Hakeem Jeffries, & Jerrold Nadler (June 8, 2020).

14. Amazon Staff, *We Are Implementing a One-Year Moratorium on Police Use of Rekognition*, AMAZON.COM, 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, 12:30 PM), <https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/>.

16. See, e.g., Olivia Solon, *Facial Recognition Bill Would Ban Use by Federal Law Enforcement*, NBC NEWS (June 25, 2020, 11:08 AM), <https://www.nbcnews.com/tech/security/2-democratic-senators-propose-ban-use-facial-recognition-federal-law-n1232128>.

17. *Id.*

18. Ethical Use of Facial Recognition Act, S. 3284, 116th Cong. § 4 (2d Sess. 2020).

19. S. 1385, 191st Gen. Ct. (Mass. 2019).

and the City of San Francisco enacted legislation prohibiting municipal departments from using FRT.²⁰

We 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 the need to understand the *accuracy* of facial recognition systems within the exact context of their intended use.²¹ The federal Facial Recognition Technology Warrant Act 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;²² the federal 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”;²³ the State of Washington requires FRT vendors to enable “legitimate, independent, and reasonable tests . . .” for “accuracy and unfair performance differences across distinct subpopulations”;²⁴ and the State of Massachusetts proposes “standards for minimum accuracy rates . . .”²⁵ 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 similarly emphasizes the need to make accuracy assessments for different population groups,²⁶ and the European Commission emphasizes the need to demonstrate robustness and accuracy with artificial intelligence (AI) systems.²⁷

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.²⁸ 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 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 FRT for a given real-world task—e.g., identifying individuals

20. S.F., Cal., Admin. Code § 19B.2. (2021).

21. *E.g.*, S. 3284 § 6(c)(1)(A)–(B).

22. Facial Recognition Technology Warrant Act of 2019, S. 2878, 116th Cong. § 5(b)(1) (1st Sess. 2019).

23. S. 3284 § 6(c)(1)(B).

24. S. 6280, 66th Leg. § 6(1)(a) (Wash. 2020).

25. Mass. S. 1385 § 14(b)(iii).

26. Rep. of Eur. Union Agency for Fundamental Rts. (FRA), *Facial Recognition Technology: Fundamental Rights Considerations in the Context of Law Enforcement* 10 (2019).

27. *Commission White Paper on Artificial Intelligence - A European Approach to Excellence and Trust*, at 20 COM (2020) 65 final (Feb. 19, 2020).

28. *See, e.g.*, JONATHAN LIPPMAN, NICHOLAS CASSIMATIS, & AARON M. RENN, CLEARVIEW AI ACCURACY TEST REPORT 1–3 (2019).

from stills of closed-circuit television video capture—is not in fact an easy undertaking. Many FRT vendors advertise stunning performance of their software.²⁹ 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 both well-lit, clear images and proper software usage from machine-learning professionals,³⁰ but during deployment, clients such as law enforcement, who may or may not have technical training, may use and evaluate data from FRT based on live video from police body cameras with very distinct image conditions.³¹ 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 computer-science literature.³²

One central concern of such cross-domain performance, which has given rise to profound criticisms of FRT, is that models may result in sharply different performance across demographic groups.³³ Models trained disproportionately on light-skinned individuals, for instance, may perform poorly on dark-skinned individuals.³⁴ For example, a leading report found that false positive rates varied by factors of 10 to 100 across

29. See, e.g., REALNETWORKS, FACIAL RECOGNITION PLATFORM 4–6 (2019); *Face Compare SDK*, FACE++, <https://www.faceplusplus.com/sdk/face-comparing/> (last visited Apr. 27, 2021); Jonathan Greig, *Air Force Hires Trueface for Facial Recognition on Bases*, TECHREPUBLIC (Nov. 19, 2019, 10:18 AM), <https://www.techrepublic.com/article/air-force-hires-trueface-for-facial-recognition-on-bases/>; *FAQs*, KAIROS, <https://www.kairos.com/faq> (last visited Apr. 27, 2021); *NEC Face Recognition Technology Ranks First in NIST Accuracy Testing*, NEC (Oct. 3, 2019), https://www.nec.com/en/press/201910/global_20191003_01.html.

30. See, e.g., PATRICK GROTH, MEI NGAN, & KAYEE HANAOKA, NAT'L INST. OF STANDARDS & TECH., ONGOING FACE RECOGNITION VENDOR TEST (FRVT) PART 1: VERIFICATION 29–30 (2019) [hereinafter VERIFICATION]; Gary Huang, Manu Rames, Tamara Berg, & Erik Learned-Miller, *Labeled Faces in the Wild Home*, U. MASS. AMHERST, <http://vis-www.cs.umass.edu/lfw/> (last updated Jan. 9, 2018, 8:15 PM).

31. Clare Garvie, *Garbage In, Garbage Out*, GEO. L. CTR. ON PRIV. & TECH. (May 16, 2019), <https://www.flawedfacedata.com>.

32. Amos Storkey, *When Training and Test Sets are Different: Characterizing Learning Transfer*, in DATASET SHIFT IN MACHINE LEARNING 3–7 (Joaquin Quiñero-Candela, Masashi Sugiyama, Anton Schwaighofer, & Neil D. Lawrence eds., 2009); VLADIMIR N. VAPNIK, THE NATURE OF STATISTICAL LEARNING THEORY 35 (2d ed. 2000); Olivier Bousquet & André Elisseeff, *Stability and Generalization*, 2 J. MACH. LEARNING RSCH. 499, 499–500 (2002); Hidetoshi Shimodaira, *Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function*, 90 J. STAT. PLAN. & INFERENCE 227, 228 (2000); ALEXEY TSYMBAL, TRINITY COLL. DUBLIN, THE PROBLEM OF CONCEPT DRIFT: DEFINITIONS AND RELATED WORK (2004). 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 & Qiang Yang, *A Survey on Transfer Learning*, 22 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENG'G 1345, 1345 (2010).

33. *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software*, NAT'L INST. STANDARDS & TECH., <https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software> (May 18, 2020).

34. PATRICK GROTH, MEI NGAN, & KAYEE HANAOKA, FACE RECOGNITION VENDOR TEST (FRVT) PART 3: DEMOGRAPHIC EFFECTS, NAT'L INST. OF STANDARDS & TECH. 1–3, 16–17, 54 (2019) [hereinafter DEMOGRAPHIC EFFECTS]; Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in 81 PROC. MACH. LEARNING RSCH. 77, 77 (2018).

demographic groups, with such errors being “highest in West and East African and East Asian people, and lowest in Eastern European individuals.”³⁵

In this Article, we characterize this gulf between the contexts in which FRT is created and the contexts in which FRT is 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 assess 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, 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 Article to that subject.

35. DEMOGRAPHIC EFFECTS, *supra* note 34, at 2.

Of course, many other considerations factor into the adoption of FRT. Concerns over privacy,³⁶ consent,³⁷ transparency,³⁸ and biased usage³⁹ 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 Article. 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 Article.

I. THE CHALLENGE OF PERFORMANCE ASSESSMENT IN NEW DOMAINS

FRT vendors report stunning accuracies of their products: SAFR,⁴⁰ Kairos,⁴¹ Face⁺⁺,⁴² and others report accuracies above 99% on National Institute of Standards and Technology (NIST) tests and other benchmark

36. 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 U.S. GOV'T ACCOUNTABILITY OFF., GAO-15-621, FACIAL RECOGNITION TECHNOLOGY: COMMERCIAL USES, PRIVACY ISSUES, AND APPLICABLE FEDERAL LAW (2015); *Seeing is ID'ing: Facial Recognition and Privacy*, CTR. 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.

37. See generally Evan Selinger & Woodrow Hartzog, *The Inconsistency of Facial Surveillance*, 66 LOY. L. REV. 101 (2019) (thorough investigation of facial recognition and consent). 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 way to go 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*, INDEPENDENT (Jan. 31, 2019, 10:19 PM), <https://www.independent.co.uk/news/uk/crime/facial-recognition-cameras-technology-london-trial-met-police-face-cover-man-fined-a8756936.html>.

38. Relatedly, public and private use of FRT to date have been handled with little transparency as several civil-society organizations, such as the Electronic Frontier Foundation (EFF), American Civil Liberties Union (ACLU), and Project on Government Oversight, were denied Freedom of Information Act (FOIA) and Right-to-Know Law (RTK) requests over what technologies were used for what purposes in different areas. See Lisa Limb & Bogyung Lim, Case Profile of *Elec. Frontier Found. v. DOJ*, UNIV. OF MICH. L. SCH.: THE C.R. LITIG. CLEARINGHOUSE (Jul. 13, 2020), <https://www.clearinghouse.net/detail.php?id=16071>; Taylor Telford, *ICE Refuses to Turn Over Internal Documents on Facial Recognition Tech and Detention Tactics, Lawsuit Says*, WASH. POST (Nov. 7, 2019, 2:56 PM), <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.

39. 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 poor people 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>.

40. REALNETWORKS, *supra* note 29, at 3–5.

41. *FAQs*, *supra* note 29.

42. *Face Compare SDK*, *supra* note 29.

datasets, such as Labeled Faces in the Wild (LFW).⁴³ 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.⁴⁴ 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.⁴⁵ 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.

A. 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.⁴⁶ 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, hairstyles, glasses, facial hair, and 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.⁴⁷ This is due to the fact that machine learning models are built to recognize patterns in a certain distribution (i.e., 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 the model learned may not apply and accuracy will likely suffer.⁴⁸ Accuracy guarantees of machine learning models depend upon the similarity of training, testing, and

43. Huang et al., *supra* note 30.

44. See, e.g., 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, 521 (2012).

45. In the machine learning literature, these data differences are more precisely described as covariate shift, prior probability shift, and concept shift. See *id.* at 522–25.

46. *Id.* at 521–22.

47. See generally Bousquet & Elisseeff, *supra* note 32, at 499; Huang et al., *supra* note 30.

48. See generally Bousquet & Elisseeff, *supra* note 32, at 499; Moreno-Torres et al., *supra* note 44, at 521.

deployment of images.⁴⁹ If images fall outside of the training or 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.⁵⁰ These datasets are commonly regarded as dissimilar to several real-life deployment domains.⁵¹ Subjects are, for the most part, aware that their pictures are intentionally being taken 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,⁵² the size of these datasets is much smaller than those scraped off the internet.⁵³ Moreover, subject skin color, hairstyle, and age 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.⁵⁴ The datasets used by FRT vendors to report accuracies, such as LFW, consist of images that are vastly different from those coming from most of the application domains where FRT is actually used.⁵⁵ The LFW creators themselves acknowledge these limitations, noting that “[n]o 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 additionally, “poor lighting, extreme pose, strong occlusions, [and] low resolution . . . do not constitute a major part of LFW” which thus disqualifies the dataset for use as a commercial benchmark.⁵⁶ While NIST has done leading work to benchmark FRT, NIST benchmark datasets are

49. VAPNIK, *supra* note 32, at 35–57; Bousquet & Elisseeff, *supra* note 32. 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 we 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 do utilize domain adaptation, there still is no substitute for training a model on the correct deployment domain.

50. 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>.

51. *Id.*; Huang et al., *supra* note 30.

52. See, e.g., Adam Harvey & Jules LaPlace, *MegaPixels: Face Recognition Training Datasets*, MEGAPIXELS (Jan. 2020), <https://ahprojects.com/megapixels/>.

53. For instance, Clearview AI claims to have scraped over three billion images from the likes of Facebook, YouTube, Venmo, and other sites. Kashmir Hill, *The Secretive Company that Might End Privacy as We Know It*, N.Y. TIMES (Mar. 18, 2021), <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>.

54. See Huang et al., *supra* note 30.

55. *Id.*

56. *Id.*

still substantially more controlled than many applications.⁵⁷ 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.⁵⁸ The rest of the images considered in that study were mug shot images of varying quality, which are also often centered on the face of a cooperative subject.⁵⁹

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

B. 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.”⁶⁰ As articulated by Ben Green and Yiling Chen,⁶¹ the performance of AI systems is often understood through statistical metrics of accuracy, but technical accuracy does not reflect the true effect the technology has in the field because humans still typically act on that technology.⁶²

Such performance differences can arise even with identical imagery by vendors and users.⁶³ 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.⁶⁴ One 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,⁶⁵ one sheriff’s office reported, for instance, “We do not set nor do we utilize a confidence threshold.”⁶⁶ Operational performance would likely be quite different for a department that abides by Amazon’s recommendations.

57. PATRICK GROTH, MEI NGAN, & KAYEE HANAOKA, FACE RECOGNITION VENDOR TEST (FRVT) PART 2: IDENTIFICATION, NAT’L INST. OF STANDARDS & TECH. 2, 11, 14–15 (2018) [hereinafter IDENTIFICATION].

58. *Id.* at 15.

59. *Id.* at 14–15.

60. Lucy Suchman, Jeanette Blomberg, Julian E. Orr, & Randall Trigg, *Reconstructing Technologies as Social Practice*, 43 AM. BEHAV. SCIENTIST 392, 399 (1999).

61. Ben Green & Yiling Chen, *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments*, in 2019 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY (FAT*) 90, 90 [hereinafter *Disparate Interactions*].

62. *Id.* at 90–91.

63. See IDENTIFICATION, *supra* note 57, at 5 (“Virtually all applications using automated face search require human review of the outputs at some frequency. . . . Human reviewers make recognition errors. . . .”).

64. See LUCAS D. INTRONA & HELEN NISSENBAUM, FACIAL RECOGNITION TECHNOLOGY: A SURVEY OF POLICY AND IMPLEMENTATION ISSUES, CTR. FOR CATASTROPHE PREPAREDNESS & RESPONSE 29–35 (2009).

65. AMAZON REKOGNITION: DEVELOPER GUIDE, AMAZON, INC. 159, <https://docs.aws.amazon.com/rekognition/latest/dg/rekognition-dg.pdf#considerations-public-safety-use-cases> (last visited Apr. 27, 2021).

66. Bryan Menegus, *Defense of Amazon’s Face Recognition Tool Undermined by Its Only Known Police Client*, GIZMODO (Jan. 31, 2019, 4:55 PM), <https://gizmodo.com/defense-of-amazons->

Much research documents the potential divergence between raw model output and human decisions based on that output.⁶⁷ Joint human–AI system breakdown can stem from several issues. Users may ignore model output, either because they do not understand or trust the system or they view themselves as more qualified, as some experiments with judges using pretrial risk assessment algorithms suggest.⁶⁸ 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).⁶⁹ Users have also been shown to selectively listen to model output that confirms their own biases, which can lead to amplified discrimination concerns.⁷⁰

Where FRT is embedded in a human system, understanding its performance also requires understanding the impact on human decision-makers. Such performance measurement would be particularly valuable to incentivize best practices (e.g., training and communication) in the adoption of FRT, rather than mandating specific practices.⁷¹

II. 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 in 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

face-recognition-tool-undermined-by-1832238149; see also JENNIFER LYNCH, FACE OFF: LAW ENFORCEMENT USE OF FACE RECOGNITION TECHNOLOGY, ELEC. FRONTIER FOUND. 15–16 (Gennie Gebhart ed. 2020).

67. See *infra* notes 69–71 and accompanying text.

68. See, e.g., Megan Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 304, 334 (2018); MEGAN T. STEVENSON & JENNIFER L. DOLEAC, THE ROADBLOCK TO REFORM, AM. CONST. SOC'Y 2–4, 6–9 (2018); Angèle Christin, *Algorithms in Practice: Comparing Web Journalism and Criminal Justice*, 4 BIG DATA & SOC'Y 1, 6, 8–9, 11 (2017).

69. See, e.g., Birte Englich, Thomas Mussweiler, & Fritz Strack, *Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making*, 32 PERSONALITY & SOC. PSYCH. BULL. 188, 188–89, 197–98 (2006); Aaron Springer, Victoria Hollis, & Steve Whittaker, *Dice in the Black Box: User Experiences with an Inscrutable Algorithm*, CORNELL UNIV.: ARXIV.ORG 1–2 (Dec. 7, 2018), <https://arxiv.org/abs/1812.03219>.

70. See, e.g., *Disparate Interactions*, *supra* note 61; Bo Cowgill, *The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities* 1–3, 7–8, 12–13 (Dec. 5, 2018) (unpublished manuscript) (on file with author); Alex Albright, *If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions* 1–4, 34 (Sept. 3, 2019) (unpublished manuscript) (on file with author); Jennifer Skeem, Nicholas Scurich, & John Monahan, *Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants*, 51 L. & HUM. BEHAV. 1, 1–6, 14–18, 21 (2020).

71. 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, 1333–35, 1339 (1985).

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.⁷² 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 addresses the data-centered roadblocks to establishing reliable in-domain accuracy. The second set of recommendations focuses on evaluating the human component affecting in-domain FRT system accuracy.

A. 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. Even if users wanted to assess domain shift, not all systems readily enable such testing.⁷³ We hence provide a protocol that would enable such assessments.

1. Vendor and Third-Party Data Transparency

a. 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 and testing data along with documentation,⁷⁴ which enables users to compare images to assess the potential for domain shift.

b. Fallback of Random Sample and Comparison Metrics

Although publishing the full imagery is the ideal solution,⁷⁵ a less desirable fallback would be the publication of a large random sample of

72. For a more ambitious proposal that proposes a new federal regulatory agency specifically for FRT, see ERIK LEARNED-MILLER, VICENTE ORDÓÑEZ, JAMIE MORGENSTERN, & JOY BUOLAMWINI, *FACIAL RECOGNITION TECHNOLOGIES IN THE WILD: A CALL FOR A FEDERAL OFFICE* 4, 23–37 (2020).

73. See James Vincent, *The Tech Industry Doesn't Have a Plan for Dealing with Bias in Facial Recognition*, *THE VERGE* (Jul. 26, 2018, 11:19 AM), <https://www.theverge.com/2018/7/26/17616290/facial-recognition-ai-bias-benchmark-test>.

74. See Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, & Kate Crawford, *Datasheets for Datasets*, CORNELL UNIV.: ARXIV.ORG 1–2 (Mar. 19, 2020), <https://arxiv.org/pdf/1803.09010.pdf>.

75. Releasing data may be infeasible due to other constraints (e.g., privacy and intellectual property). See Andrew W. Senior & Sharathchandra Pankanti, *Privacy Protection and Face Recognition*, in *HANDBOOK OF FACE RECOGNITION* 671 (Stan Z. Li & Anil K. Jain eds. 2011); Olivia Solon, *Facial Recognition's 'Dirty Little Secret': Millions of Online Photos Scraped Without Consent*, *NBC NEWS* (May 27, 2021, 8:09 PM), <https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921>.

imagery⁷⁶ and enabling the comprehensive calculation within the 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).⁷⁷ 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.

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.

a. Enabling Testing

Vendors and users should facilitate independent validation of in-domain performance with programmatic access within their platforms. 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.

b. 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-labeled data. In addition, vendors should provide programmatic access (e.g., via API) to enable users to assess performance with

76. 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.

77. Commonly used measures for domain discrepancy include measures like maximum mean discrepancy (MMD). See, e.g., Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, & Alexander Smola, *A Kernel Two-Sample Test*, 13 J. MACH. LEARNING RSCH. 723, 725 (2012); Wouter M. Kouw & Marco Loog, *An Introduction to Domain Adaptation and Transfer Learning*, CORNELL UNIV.: ARXIV.ORG 18 (Jan. 14, 2019), <https://arxiv.org/abs/1812.11806>; Gregory Ditzler & Robi Polikar, *Hellinger Distance Based Drift Detection for Nonstationary Environments*, in 2011 IEEE SYMP. ON COMPUTATIONAL INTEL. IN DYNAMIC AND UNCERTAIN ENVT'S 41, 41–42, 48 (2011). There are several challenges to address in applying such metrics: deciding upon what aspects of the data the metric focuses on (e.g., Kullback Leibler divergence can be calculated over image brightness, color, 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. *Id.*

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. 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 time⁷⁸ or vendor software updates.⁷⁹

a. Documentation

Vendors should provide detailed release notes and documentation for each version of the FRT system, including changes in the model and data. We recommend the changes 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 may trigger user recertification.⁸⁰

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

B. Human-Centered Recommendations

While evaluating in-domain accuracy is necessary to understand operational performance, technical accuracy alone remains inadequate. FRT outputs are used by humans, deployed within existing institutions, and interact in a social setting.⁸¹ Understanding the human-FRT interaction is 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.”⁸² Even though most computer vision research focuses on optimizing accuracy isolated from its human surroundings, we urge an assessment that encompasses accuracy in context.⁸³

78. TSYMBAL, *supra* note 32, at 2.

79. See, e.g., Alessandro Mascellino, *iOS Update Introduces Facial Recognition with Masks to iPhones, but Only Via Apple Watch*, BIOMETRIC (Feb. 3, 2021), <https://www.biometricupdate.com/202102/ios-update-introduces-facial-recognition-with-masks-to-iphones-but-only-via-apple-watch>.

80. We note that these recommendations are similar in spirit to the certification ideas in LEARNED-MILLER ET AL., *supra* note 72, at 20–22.

81. See, e.g., Ben Green & Yiling Chen, *The Principles and Limits of Algorithm-in-the-Loop Decision Making*, in 3 PROC. ACM HUM.-COMPUT. INTERACTION 1, 1–2 (2019) [hereinafter *Principles and Limits*].

82. *Id.* at 4.

83. Bryan Wilder, Eric Horvitz, & Ece Kamar, *Learning to Complement Humans*, CORNELL UNIV.: ARXIV.ORG 1–2 (May 1, 2020), <https://arxiv.org/abs/2005.00582>.

The most direct approach to test FRT's in-domain operational accuracy 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 was randomly disclosed to police districts and outcomes were compared across districts.⁸⁴ The trial found no statistically significant evidence of crime reduction from predictive policing.⁸⁵ 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.⁸⁶ That said, such end-to-end field experiments can be resource-intensive to design and administer⁸⁷ and pose potential ethical challenges and risks to a community.⁸⁸ Moreover, in the FRT context, there are significant questions about which objective outcome measures are available. In the police identification context, for instance, arrests and convictions may be affected by bias⁸⁹ or perceptions of FRT.⁹⁰ 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 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."⁹¹ In certain enforcement contexts, FRT face

84. PRISCILLIA HUNT, JESSICA SAUNDERS, & JOHN S. HOLLYWOOD, EVALUATION OF THE SHREVEPORT PREDICTIVE POLICING EXPERIMENT 4–27 (2014). *See generally* Jessica Lee, Yanwen Lin, & Michael A. Madaio, *A Longitudinal Evaluation of a Deployed Predictive Model of Fire Risk*, in 32D CONF. NEURAL INFO. PROCESSING SYS. (2018) (describing an in-domain assessment of fire prediction software).

85. HUNT ET AL., *supra* note 84, at 49–50.

86. ALAN S. GERBER & DONALD GREEN, FIELD EXPERIMENTS: DESIGN, ANALYSIS, AND INTERPRETATION 8–15 (Ann Shin, Jake Schindel, & Jack Borrebach eds., 2012).

87. *But see* Cassandra Handan-Nader, Daniel E. Ho, & Becky Elias, *Feasible Policy Evaluation by Design: A Randomized Synthetic Stepped-Wedge Trial of Mandated Disclosure in King County*, 44 EVALUATION REV. 3, 4 (2020).

88. *See, e.g.*, Kim Yi Dionne, Augustine Harawa, & Hastings Honde, *The Ethics of Exclusion When Experimenting in Impoverished Settings*, in ETHICS AND EXPERIMENTS: PROBLEMS AND SOLUTIONS FOR SOCIAL SCIENTISTS AND POLICY PROFESSIONALS 25–40 (Scott Desposato ed. 2016).

89. *See* Andrew Gelman, Jeffrey Fagan, & Alex Kiss, *An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias*, 102 J. AM. STAT. ASS'N 813, 813–14, 821–22 (2007).

90. *See* Ajay Sandhu & Peter Fussey, *The 'Uberization of Policing'? How Police Negotiate and Operationalise Predictive Policing Technology*, 31 POLICING & SOC'Y 66, 75–77 (2021).

91. Linda J. Skitka, Kathleen L. Mosier, & Mark Burdick, *Does Automation Bias Decision-Making?*, 51 INT'L J. HUM.-COMPUT. STUD. 991, 992–94, 999, 1002–03 (1999); Jeffrey Warshaw,

matches are trusted without regard to reported system accuracy or confidence of output.⁹² 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.”⁹³ In the criminal justice context, for instance, some judges entirely ignore risk assessment scores.⁹⁴ 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 with moderate risk scores than white defendants with moderate risk scores.⁹⁵

We recommend, wherever possible, A/B testing to assess the impacts of the FRT output on specific human decisions. When an FRT system delivers model output to human decision-makers, 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 line-ups, 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 overreliance 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 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 preexisting (non-FRT) system. A comparison between the manual and FRT-augmented

Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, & Barton A. Smith, *Can an Algorithm Know the “Real You”?* *Understanding People’s Reactions to Hyper-personal Analytics Systems*, in PROC. 33D ANN. ACM CONF. ON HUM. FACTORS IN COMPUTING SYS. 797, 802–05 (2015); see also Englich et al., *supra* note 69, at 197; Springer et al., *supra* note 69, at 3–4.

92. See Garvie, *supra* note 31; Menegus, *supra* note 66.

93. Berkeley J. Dietvorst, Joseph P. Simmons, & Cade Massey, *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, 144 J. EXPERIMENTAL PSYCH.: GEN. 114, 114–15 (2014).

94. Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. 439, 492 (2020).

95. Albright, *supra* note 70, at 4.

decisions would enable an assessment of the effects of the FRT system on performance.⁹⁶

We recognize 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 or budget) cannot easily be assessed without many A/B tests across institutions. But the more A/B 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.

III. RESPONSIBILITIES BEYOND THE PROTOCOL

We now spell out several other recommendations for implementation of in-domain testing. Users and vendors have particular responsibilities in adopting and implementing the protocol. 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.

A. Service Model

The recommendations of this Article may ultimately not be achievable through the 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 ongoing 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.

B. 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 Article and demand evidence for performance in the user's specific domain.

96. See David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 *YALE J. ON REG.* 800, 849–53 (2020).

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.⁹⁷ Similarly, a pilot could compare different vendors as the New York Police Department did during a trial period with predictive policing.⁹⁸ 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.

C. 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).⁹⁹ Over the past twenty years, FRVT has benchmarked algorithms for performance in facial identification, facial verification, and other tasks.¹⁰⁰ Vendors submit an executable version of their algorithm to NIST, which NIST deploys on its non-public datasets to compute the algorithm's accuracy.¹⁰¹ While NIST's benchmarking is exemplary,¹⁰² 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.

97. See U.S. CUSTOMS & BORDER PROT., SOUTHERN BORDER PEDESTRIAN FIELD TEST: SUMMARY REPORT (2016).

98. N.Y. POLICE DEP'T, PREDICTIVE POLICING PILOT EVALUATION (June 2016), <https://www.brennancenter.org/sites/default/files/Predictive%20Policing%20Final%20802-818%20-%20%28%23%20Legal%208799488%29.PDF>. Note that this document was obtained through a Freedom of Information Law (FOIL) request by the Brennan Center for Justice. See *NYPD Predictive Policing Documents*, BRENNAN CTR. FOR JUST. (July 12, 2019), <https://www.brennancenter.org/our-work/research-reports/nypd-predictive-policing-documents>.

99. Lauren Chambers, *Five Fast Facts from the Federal Study of Demographic Bias in Facial Recognition*, PRIV. SOS (Feb. 3, 2020), <https://privacysos.org/blog/five-fast-facts-from-the-federal-study-of-demographic-bias-in-facial-recognition/> (“The FRVT is widely regarded as an international standard for facial recognition.”); *NEC's Face Recognition Technology Ranks First in NIST Testing for Third Consecutive Time*, BUS. WIRE (July 15, 2014, 9:59 AM), <https://www.business-wire.com/news/home/20140715006057/en/NECs-Face-Recognition-Technology-Ranks-First-in-NIST-Testing-for-Third-Consecutive-Time> (“The NIST report is widely viewed as the benchmark of determining which face recognition software is the most accurate. . .”).

100. *Face Recognition Vendor Test (FRVT)*, NIST: PROJECTS/PROGRAMS, <https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt> (last visited Apr. 27, 2021); see, e.g., VERIFICATION, *supra* note 30; IDENTIFICATION, *supra* note 57; DEMOGRAPHIC EFFECTS, *supra* note 34.

101. Vincent, *supra* note 73.

102. See *id.* (quoting Clare Garvie) (“NIST does a very admirable job in conducting these tests[.]”).

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.¹⁰³

D. Academics

While academics have played a critical role in surfacing errors and biases in a range of AI systems—including recidivism prediction,¹⁰⁴ predictive policing,¹⁰⁵ medical diagnostics,¹⁰⁶ and FRT systems¹⁰⁷—and spearheading basic FRT research,¹⁰⁸ 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 protocol should enable academic researchers to engage in more rigorous assessments—as third-party evaluators in collaboration with FRT users—of the performance across uncharted domains.

E. 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 contexts¹⁰⁹ and FRT in particular.¹¹⁰ 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.

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,

103. See Jodi L. Short & Michael W. Toffel, *The Integrity of Private Third-Party Compliance Monitoring*, 42 ADMIN. & REGUL. L. NEWS 22, 22–24 (2016).

104. See, e.g., *Disparate Interactions*, *supra* note 61, at 2–5; Cowgill, *supra* note 71, at 2, 11–12; Albright, *supra* note 70, at 2, 25–30, 34.

105. Kristian Lum & William Isaac, *To Predict and Serve?*, 13 SIGNIFICANCE MAG. 14, 16–19 (2016).

106. Ziad Obermeyer, Brian Powers, Christine Vogeli, & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 449–50, 452–53 (2019).

107. Buolamwini & Gebru, *supra* note 34, at 86–88.

108. See, e.g., Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, & Yi Yang, *A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification*, in 33D PROC. OF THE AAAI CONF. ON A.I. 8738, 8738–39, 8742–44 (2019).

109. See, e.g., Julia Angwin, Jeff Larson, Surya Mattu, & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

110. See, e.g., Jacob Snow, *Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots*, ACLU (July 26, 2018, 8:00 AM), www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

reflection, and consideration of one of the most important emerging applications of AI today.

APPENDIX: SUMMARY OF RECOMMENDATIONS

Our recommendations focus on two sources of performance differences between the development and deployment contexts: data-centered recommendations and human-centered recommendations.

A. Data-Centered Recommendations

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

1. Protocol for Domain-Specific Model Assessment

a. 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.

b. 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.

c. 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.

B. Human-Centered Recommendations

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).

1. 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, 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 indicate the confidence of identification, for a random subset of images may enable an assessment of overreliance and the potential for selective reliance.

2. Procurement and Auditing

a. 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 (1) technical accuracy and (2) the effects of FRT on human decision-making.

b. Expanding Auditing

With the protocol 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.