Stanford University
Human-Centered
Artificial Intelligence

NOVEMBER 2020

Domain Shift and Emerging Questions in Facial Recognition Technology

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

FACIAL RECOGNITION TECHNOLOGIES HAVE GROWN in sophistication and adoption throughout American society. Consumers now use facial recognition technologies (FRT) to unlock their smartphones and cars; retailers use them for targeted advertising and to monitor stores for shoplifters; and, most controversially, law enforcement agencies have turned to FRT to identify suspects. Significant anxieties around the technology have emerged—including privacy concerns, worries about surveillance in both public and private settings, and the perpetuation of racial bias.

In January 2020, Detroit resident Robert Julian-Borchak Williams was wrongfully arrested, in what the *New York Times* named as possibly the first instance of an arrest based on a faulty FRT algorithm. The incident highlights the role of FRT in the nation's ongoing conversation around racial injustice. The killings of George Floyd, Breonna Taylor, and Ahmaud Arbery and the public demonstrations that followed in the spring and summer of 2020 compelled a long overdue reckoning with racial injustice in the United States. FRT systems have been documented to exhibit worse performance with darker-skinned individuals and we must hence examine the potential for such technology to perpetuate existing injustices. This brief points towards an evaluative framework to benchmark whether FRT works as billed. In the face of calls for a ban or moratorium on government and police use of FRT systems, we embrace the demand for a pause so that the technical and human elements at play can be more deeply understood and so that standards for a more rigorous evaluation of FRT can be developed.

KEY TAKEAWAYS

- FRT vendors and developers should ensure their models are created in a way that is as transparent as possible, capable of being validated by the user, and well documented. The effect these systems have on the decision making of their users must be understood more deeply and policymakers should embrace A/B testing as a tool to gauge this.
- Users in government and business settings should condition the procurement of FRT systems on in-domain testing and adherence to established protocols.
- We support calls for a moratorium on FRT adoption in government and policing while a more responsible testing framework is developed.

Our recommendations in this brief extend to both the computational and human side of FRT. In seeking to answer how we bridge the gap between testing FRT algorithms in the lab and testing products under real world conditions, we focus on two sources of uncertainty: first, the specific differences in model output between development settings and end user applications (which we term here *domain shift*), and second, the differences in end user interpretation and usage of model output across the institutions employing FRT (which we refer to as *institutional shift*). Policymakers have a crucial role to play in ensuring that responsible protocols for FRT assessment are codified—both as they pertain to the impact FRT have on human decision making as well as how they pertain to the performance of the technology itself. In building out a framework for responsible testing and development, policymakers should further look to empowering regulators to use stronger auditing authority and the procurement process to prevent FRTs from evolving in ways that would be harmful to the broader public.

Introduction

In May 2020, we hosted a workshop to discuss the performance of facial recognition technologies that included leading computer scientists, legal scholars, and representatives from industry, government, and civil society. The white paper this workshop produced, "Evaluating Facial Recognition Technology: A Protocol for Performance Assessment in New Domains," seeks to answer key questions in improving our understanding of this rapidly changing space. While the workshop was held before the nationwide upheaval in the wake of the killings of George Floyd, Breonna Taylor, and Ahmaud Arbery, our recommendations are particularly important as nearly all proposed legislation or regulation of FRT calls for evaluation of its operational performance.

The rapid adoption of FRT across industries and complex ethical concerns about FRT's impact on society require much more substantial testing than exists

The Challenge of FRT in the Wild

The FRT industry in the United States is growing at a tremendous pace. Currently valued at $5 billion, the market for FRT systems is projected to double by 2025. While the National Institute of Standards and Technology has established the well-known FRVT (Facial Recognition Vendor Test) benchmarking standard, the rapid adoption of FRT across industries and complex ethical concerns about FRT's impact on society require much more substantial testing than exists. Many vendors currently advertise high performance metrics for their software, but these tests are carried out in the confines of carefully calibrated testing settings. Evaluating the performance of FRT for a real-world task like identifying individuals from stills of closed-circuit television in real time is a significantly more complicated task. The context in which accuracy is tested is often vastly different from the context in which the actual program is applied. For example, vendors may train their images with clear, well-lit images but during deployment law enforcement officers

> ***Domain shift** can significantly degrade model performance. Domain shift also encompasses the profound problem of bias: algorithms trained on one demographic group may perform poorly on another*

may use FRT based on live footage from police body cameras. Computer science research has established that this **"domain shift"** can significantly degrade model performance. Domain shift also encompasses the profound problem of bias: algorithms trained on one demographic group may perform poorly on another. One leading report found that false positive rates varied by factors of 10 to 100 across demographic groups, with the report citing such errors being "highest in West and East African and East Asian people, and lowest in Eastern European individuals."

In addition to these data considerations, how humans incorporate algorithmic output can also contribute to the failure of FRT systems. This **"institutional shift"** comes from the fact that the same system may be utilized in sharply different ways by companies or agencies. This type of uncertainty can stem from users selectively listening to model output that confirms their own preexisting biases, users ignoring model output, or users over-trusting an algorithm. For instance, two police departments in neighboring jurisdictions deploying identical systems could reach sharply divergent conclusions about the same model output in a suspect identification use case if one department insists on using a higher confidence threshold compared to its neighbor department. What technologists would see as accurate may be interpreted quite differently by the operator using FRT algorithms in the field.

With these two overarching sources of uncertainty in mind, we articulate recommendations for a responsible testing protocol to address these challenges.

Policy Discussion

To address negative outcomes stemming from domain-specific concerns, we recommend policymakers focus on the following three pieces of a larger testing protocol. First, vendors and developers should put greater emphasis on transparency in their training data. Ideally, this would consist of the full vendor training and test set imagery being made available to the public. If this is not feasible, an alternative regime could use large random samples of imagery to facilitate comparative analysis of any discrepancies between vendor and user metrics.

Second, vendors should provide users and third-party evaluators meaningful access to testing imagery so that they can conduct independent validation of in-domain performance. Such access should also allow users to label their own testing data, reserve holdout testing data, and define metrics that must be met prior to commercial deployment.

Third, vendors and users should conduct ongoing, periodic recertifications of FRT performance and vendors should provide comprehensive release notes and documentation for each version of the model in

Vendors should provide users and third-party evaluators meaningful access to testing imagery so that they can conduct independent validation of in-domain performance

question. These release notes should, at minimum, include changes to the underlying model, performance metrics across subcategories like demographics and image quality, and potentially be used to trigger a recertification process if one becomes necessary.

To address FRT performance issues stemming from human decision making, policymakers should encourage A/B testing to assess performance within the human context. This would enable researchers to evaluate the effect an FRT system on human decision making. A/B testing can be adapted to compare human decisions with AI-augmented decisions, to assess the human operator's responsiveness to "confidence scores" of models, or to gauge potential over-reliance or under-reliance on model output (sometimes referred to as "automation bias" and "algorithm aversion," respectively).

Opening up facial recognition systems to facilitate in-domain accuracy testing will empower a much wider range of parties and stakeholders to rigorously assess the technology. By following the protocols spelled out here, watchdog organizations can expand performance benchmarking on a standardized basis and audit systems on a wider range of FRT domains more rapidly. Businesses and government agencies procuring facial recognition systems through large-scale contracts should condition such purchases on rigorous in-domain accuracy tests adhering to the evaluative framework articulated above. Auditors should expand their testing datasets to cover high-priority emerging domains and academic researchers should pursue more research on domain drift in FRT. Finally, media and civil society organizations should amplify the findings of this new testing framework to ensure FRT is better understood in public settings. While a moratorium on facial recognition technologies in criminal justice is laudable step at this time, FRT may continue to be deployed across settings and standards for whether and how to adopt FRT must be worked through now. Adopting these protocols and recommendations will not—and should not—silence legitimate scrutiny of facial recognition technology, but our hope is that providing a conceptual framework here to evaluate/test the negative effects of domain shift and institutional shift can offer a crucial next step for better understanding the operational and human impacts of this emerging technology.

Policymakers should encourage A/B testing to assess performance within the human context

The original White Paper, **"Evaluating Facial Recognition Technology: A Protocol for Performance Assessment in New Domains,"** can be found at https://hai.stanford.edu/sites/default/files/2020-11/HAI_FacialRecognitionWhitePaper.pdf

Stanford Institute for Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Daniel E. Ho is a William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, Senior Fellow at the Stanford Institute for Economic Policy Research, Stanford University; and Associate Director, Stanford Institute for Human-Centered Artificial Intelligence.

Emily Black is a Ph.D. Student, Computer Science Department, Carnegie Mellon University.

Maneesh Agrawala is a Forest Baskett Professor of Computer Science and Director of the Brown Institute for Media Innovation at Stanford University, and Affiliated Faculty, Stanford Institute for Human-Centered Artificial Intelligence.

Fei-Fei Li is a Sequoia Professor in the Computer Science Department at Stanford University, and Co-Director, Stanford Institute for Human-Centered Artificial Intelligence.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**