FACIAL RECOGNITION TECHNOLOGIES IN THE WILD:

A CALL FOR A FEDERAL OFFICE

Erik Learned-Miller, Vicente Ordóñez, Jamie Morgenstern, and Joy Buolamwini

May 29, 2020

Acknowledgements:
Facial Recognition Technologies in the Wild: A Call for a Federal Office and the supplemental document *Facial Recognition Technologies: A Primer* were made possible by a grant from the John D. and Catherine T. MacArthur Foundation to the Algorithmic Justice League.

We thank the following individuals for comprehensive feedback on earlier drafts of these documents: Sasha Costanza-Chock, Aaina Agarwal, Ben Hutchinson, Brant Cheikes, David Evans, and Paul Humphreys.

ALGORITHMIC JUSTICE LEAGUE

MacArthur Foundation

Contents

SECTION 1

Introduction

In recent years, facial recognition technologies (FRTs) have experienced enormous growth and rapid deployment.[1] The potential benefits of FRTs such as increased efficiency, diagnosis of medical conditions, and the ability to find persons of interest are tempered with risks of mass surveillance, disparate impact on vulnerable groups, algorithmic bias, and lack of affirmative consent.

The passage of city and statewide restrictions [9, 8, 10, 7] and proposed federal legislation [58, 22, 23, 21] show growing public concern. They also demonstrate the need for comprehensive policies to address the wide range of uses across private and public sectors. Current legislative efforts address a patchwork of different applications, jurisdictions, and time periods. They do not cover the full scope and spread of FRTs.

The ubiquitous scenarios that lawmakers have not yet addressed require oversight and guidance for industry practice, research norms, procurement procedures, and categorical bans where deemed appropriate. Depending upon the application, societal, legal, ethical, financial and even physical risks demand a thorough understanding of real-world impacts. How can we manage such a complex set of technologies with such enormous societal implications?

Many other authors have addressed the ethical and societal implications of FRTs [65, 11, 47, 69] and of artificial intelligence more broadly [24, 40, 4]. Several groups have argued for new laws and regulation of face recognition technologies (see, for example, [67, 20]). These earlier works have illuminated the widespread problems that emerge with the deployment of FRTs and related technologies. Yet in isolation they are not enough. It is time to take the next step and make a specific proposal about how to move forward. This white paper makes the following central claim.

Central Claim

Addressing the trade-offs among the risks and benefits of complex facial recognition technologies requires the creation of a new federal office.

We do not make this claim lightly. As professionals who have been involved with FRTs and adjacent technologies in both academia and industry, we have been at the center of many discussions about how to address the wide range of challenges posed by FRTs. A sampling of proposed remedies include:

- building better face databases for development and testing that display more diversity across parameters such as race, gender and age;

[1] Borrowing from the Federal Trade Commission [1], we use the term "facial recognition technologies" as a catchall phrase to describe a set of technologies that process imaging data to perform a range of tasks on human faces, including detecting a face, identifying a unique individual, and estimating demographic attributes.

- building better unified national testing standards for FRTs to improve assurances that technology is ready for deployment;
- proposals for ethical standards and principles for the development of FRTs;
- legislation to address issues of privacy and data ownership, such as the European General Data Protection Regulation (GDPR); and
- legislation to ban or restrict certain uses of FRTs, some of which have already been proposed and enacted across US cities and states.

We understand the rationale for these ideas and believe they contain important elements of a more complete solution. However, we will argue here that these approaches are not enough without coordination.

1.1 A model for the regulation of FRTs

We present our rationale for a new federal office by examining how other complex technologies have been successfully managed at the federal level. Specifically, we draw analogies with regulatory structures for two other complex industries–the medical device industry and the pharmaceutical industry. We argue that FRT raises similar questions and concerns, and has a similar potential for successful regulation through such mechanisms. Furthermore, without such mechanisms, current problems are likely to persist.

The US Food and Drug Administration (FDA), in collaboration with industry, lawmakers, and the professional medical community, has developed an extensive set of definitions, procedures, policies, conventions, laws, and regulations that have successfully managed trade-offs among the many parties involved in this domain. We assert that similar structures are required for FRTs.

To make this case, we carefully examine the risks and implications of FRTs. Then, through analogies with FDA regulatory structures, we propose specific methodologies for managing the risk- benefit trade-offs of this technology. This includes definitions to simplify and clarify key concepts, the classification of applications into risk categories, the adoption of scoping concepts like “intended use”, and recommendations about appropriate gatekeepers for different parts of the process. We address issues around deployment restrictions, research and development, consent and privacy, training of users, and mandatory reporting of errors. Addressing the entire “FRTs ecosystem” as a whole—from developers, to users, to impacted populations and regulators—is necessary to understand the trade-offs and forces at work with these technologies. While theoretically these components could be created and managed through other means, such as a self-regulated industry consortium, business priorities will not always align with public interest priorities. As such a federal office is the more appropriate mechanism.

This is just the beginning of a long and complex conversation needed to evaluate such a proposal. We do not pretend to have all of the answers about how such an office should be created. In fact, we remain agnostic about the precise positioning of such an office within the federal government. But this does not prevent us from discussing the elements that are necessary to establish sorely needed controls for FRTs.

Figure 1: Examples of issues with FRTs.

This figure illustrates just a few of the many potential failure modes and risks presented by FRTs. Software may overestimate its own accuracy due to the specific setting in which it is used ('Performance Assessment'). It may be used in a setting in which it was not intended ('Violation of Intended Use'). It may work poorly for certain subgroups ('Population Modeling') or may violate people's privacy in a manner that they are completely unaware of ('Privacy Violation'). These are just a few of the concerns around the use of these technologies.

1.2 Background and audience

People come to the debates on FRTs with many different backgrounds. In order to provide a common starting point, we have provided *Facial Recognition Technologies: A Primer* as an accompanying document. The primer is written for a non-technical audience and provides definitions for many of the terms used in industry and research such as *face detection, face verification*, and *face identification*. It enumerates some common and emerging applications in areas such as consumer products, schools, and police departments. And it provides a basic introduction to some of the terms used in developing and evaluating face recognition systems such as *similarity score*s, *false positives*, and *false negatives*. The primer aims to provide more details about facial recognition technologies than most material aimed at a non-technical audience, allowing for a more detailed discussion of potential benefits and harms of different use cases.

No further technical expertise in facial recognition technologies is needed to read this document. Our aim is to provide common understanding and language for a variety of stakeholders, including policy makers, procurement officers, and the broader public seeking accessible information and recommendations.

1.3 Organization of the white paper

The remainder of this document is organized as follows. In Section 2, we enumerate a wide variety of challenges that arise in the development, deployment, and management of FRTs. While many of these challenges are technical, others are ethical and sociological in nature. Still others are legal questions about the relationship among existing laws, proposed laws, and emerging technologies. In Section 3, we briefly review some proposals that have been made to address these challenges. We argue that these previous proposals are not sufficient, and that an office where these challenges can be addressed in a comprehensive and coordinated way is needed. We propose the US Food and Drug Administration as a model for the management and regulation of complex technologies. The medical device industry and pharmaceutical industry each have many parallels to the emerging industry of FRTs, and we will argue that it is logical to adopt a similar solution. We make a number of recommendations including classifying applications into risk categories and, for higher risk categories, requiring manufacturers to report statistics on errors and harms.

Finally, in Section 4, we lay out some of the core concepts that we believe need development for FRT management. Many of these concepts, such as *intended use* have strong parallels in FDA-regulated industries. We present details of a sample application, screening for hyperthyroidism using facial recognition technology, to explore some of the practical issues and real-world challenges that accompany applications of FRTs.

Many issues discussed here have been raised for the larger set of artificial intelligence (AI) technologies. Why not address these at the same time? While the issues we raise here are relevant for other technologies, FRTs are already sufficiently complex that they deserve focused consideration. We do believe that our approach could be a model for other more general artificial intelligence technologies, but we defer that discussion to a later date.

SECTION 2

Motivations

Facial recognition technologies (FRTs) have evolved rapidly over the last few decades. Yet technical challenges remain, legal uncertainties abound, and societal implications that extend beyond questions of technical accuracy or legal justification persist. These challenges must be addressed in a comprehensive manner. Existing approaches that focus on guiding principles and rely primarily on self-regulation, technical benchmarks, or legislation are insufficient. This section provides an overview of key societal, technical, and legal challenges that arise in the deployment of FRTs and motivate the call for a federal office.

2.1 Societal challenges: Principles are not enough

Principles established to guide decision making when deploying FRTs can be a starting point for developers. But external accountability and review are necessary to verify manufacturer claims and attend to societal challenges. While the motivations for FRTs can be well intended, these systems can propagate harmful discrimination, invade privacy, and rely on problematic data practices. Furthermore, there is a dangerous perception that these technologies are neutral (and hence should be given more authority than human decision makers) because they involve decisions made by machines. But recent research shows that gender, racial, and skin color biases can be propagated by commercial FRTs [16, 59, 19, 60]. Practices around disclosure, consent, and capacity for abuse or misuse pose additional challenges [28, 26, 27].

There are a variety of challenges relating to marginalized groups. One is that categories such as race and gender used in FRTs are social constructs. They are shaped by historical and cultural factors which vary across regions and change over time [63]. Classifications that rely on such social constructs and that do not have universally agreed upon definitions can be reinforced by FRTs. These classifications can limit understanding of how FRTs perform across groups that are not accounted for by commonly used classification systems. To further complicate matters, the use of FRTs on marginalized groups can increase their exposure to machine-based discrimination [63].

There is active social resistance to specific uses of FRTs. Successful grassroots campaigns have halted the planned adoption of FRTs at concerts, university campuses, and housing complexes [61, 54, 25]. Researchers who make claims about using FRTs to predict future behavior like committing a crime [73, 76] or to categorize complex traits like sexual orientation [75] face heavy scrutiny [2, 6]. In light of these important issues, we address four sets of societal challenges that need to be examined when evaluating potential deployments of FRTs: 1) consent and disclosure, 2) privacy and surveillance, 3) demographic targeting and discrimination, and 4) attacks and misuse.

Consent and disclosure. FRTs can be quickly deployed using face data available online and inexpensive camera systems. Such easy deployment enables the mass collection of personal information without consent. For systems that include consent as a design consideration, when, where, and how consent is requested is critical. For example, consider two systems which provide

an opt-out: one prior to including an individual in a dataset and the other afterwards. The latter system does not provide *affirmative consent*, the requirement of a user action before inclusion, though it is certainly better than systems without any consent mechanisms.

The lack of disclosure about the details or the existence of a system can impede due process—individuals cannot protest or seek redress to harm for decisions that are informed by FRTs they do not know about. Requirements for disclosing the intended use, capabilities, and limitations of FRTs for applications that require consent are needed to inform decisions about permissible use.

Privacy and surveillance. FRTs may be adopted in an attempt to enhance security in private areas or enable surveillance of public places to deter and detect harmful behavior, crimes, and disturbances. But deploying face recognition systems on video surveillance networks can enable mass surveillance that erodes the ability to be anonymous in a public space. Real-time persistent face surveillance violates any reasonable guarantee of privacy. Mass surveillance can also deter people from exercising rights like protesting the government or freely associating with others [3, 26]. Studies that indicate racial, gender, and age bias in face recognition systems raise concerns that the harms of misidentification fall disproportionately on already marginalized and vulnerable populations [30, 19]. Guidelines and restrictions that minimize the risk of privacy breaches and unwarranted surveillance must be established in a manner that can adapt to a wide range of public and private sector uses.

Demographic targeting and discrimination. Another area that intersects public and private sector applications is the use of FRTs to create demographic and biometric profiles of individuals based on facial features. Though the information may not reveal an individual's identity, it can allow for intentional targeting. Demographic targeting can be used in contexts including advertising and marketing as well as in law enforcement. In commerce, such practices can result in price gouging or targeted promotions that systematically exclude or exploit certain groups. Such behavior can create conditions where particular groups face persistent unfair treatment. In law enforcement, FRTs can be used to enhance harmful racial profiling [41]. Finally, the labels used by these systems can entrench social categories like race and binary gender while erasing the existence of other identities and expressions.

Beyond demographic labels, some facial recognition technologies have been developed to assign labels related to categories such as sexual orientation [75] or so-called criminality (the likelihood that someone will commit a crime) [73, 76]. Charges that someone is likely to commit a crime or assumptions about sexual orientation can pose real-world consequences for those who are labeled.[2]

[2] FRTs that attempt to somehow categorize sexual orientation or the likelihood that someone will commit a crime reduce extremely complex topics rooted in socioeconomic, historical, and structural factors to simple labels based on facial features. The classification of criminality is based on political and legal categorizations of behaviors that change over time [36], and these behaviors are not intrinsic to an individual face. Yet being labeled a criminal for any reason, let alone based solely on facial characteristics, can subject an individual to undue scrutiny and stigma at best and potential fatal interactions at worst [62]. Sexual orientation includes dimensions of identity, behavior, attraction, and arousal [12]. These dimensions are not necessarily fixed and some reflect psychological aspects that cannot (like emotions) be reliably inferred from a face. Regardless of sexual orientation being more than physical action, as of March 2019, there are 70 UN Member States that criminalize consensual same-sex sexual acts with 6 imposing the death penalty [48]. FRTs that claim to assess sexual orientation can be employed in a manner that increases harms to individuals who are labeled as homosexuals.

Furthermore, labels established for one task can be used out of context in another. For example, facial recognition technologies can attempt to analyze expressions like smiles and frowns, but claim to report the emotional state of a person, which is distinctly different [13]. When facial expression analysis is used to evaluate job candidates, cultural biases can result in harmful discrimination for qualified candidates whose facial movements are judged unfavorably by a machine [18]. FRTs that attempt to apply categorical markers to individuals need to be scrutinized for the appropriateness (do we feel comfortable inferring this information based on facial characteristics?) and reliability (how accurately can we infer this information based on facial characteristics?) of such labels.

Attacks and misuse. Finally, FRTs can rely on large stores of valuable personal data and biometric information, making these systems the target of data theft attacks.[3] Companies can also store large sets of publicly available face images in violation of the policies of the web platforms that host these images [55, 57]. Beyond external threats, internal abuse and misuse of FRTs can undermine data integrity and public trust. Currently, operators are at liberty to use systems in ways they desire, whether or not those align with the intent of the designer [27]. This opens up doors to privacy violations and exposes implicated individuals to secondary harms. Such privacy risks are an inherent danger for any system that collects large stores of valuable data. Passwords or credit cards compromised in a data breach can be changed or replaced, but faces cannot. These examples illustrate some of the societal challenges posed by FRTs which need to be examined when evaluating potential guidelines and restrictions.

2.2 Technical challenges: Benchmarks will always fall short

To mitigate issues of bias and representation in the performance of FRTs, calls to establish more comprehensive technical standards and benchmarks have been raised by researchers, policy makers, and industry leaders. While these efforts may improve the assessment of FRTs under certain conditions, they cannot solve many of the problems inherent to the use of these systems. In particular, this approach is insufficient for making determinations about permissible use of FRTs, as information about how these systems work needs to be coupled with information about where and how these systems will be deployed. Furthermore, while FRTs have seen recent advances in their technical performance, a number of key technical factors remain, including the ability to capture target application conditions, challenges with measuring performance, collecting and using benchmark data responsibly, and the difficulty of interpreting benchmark results. Appendix A presents a deeper dive into key issues of existing benchmark evaluations for FRTs. Here, we present a brief overview of benchmark limitations.

The primary issue with relying on benchmarks to inform the use of FRTs is that they can only indicate how FRTs will work in conditions that reflect the benchmark data. First consider the population represented by a benchmark. If the benchmark data is less demographically diverse than the target population (i.e, it has few or no examples of a certain subpopulation), its performance on the underrepresented groups cannot be accurately assessed. The converse is also true. If the target population is more *homogeneous* (with less variation) than the benchmark,

[3] Organizations that use FRTs often rely on vendors that can introduce additional risks. In June 2019, the US Customs and Border Protection agency confirmed that tens of thousands of images of drivers in their cars and license plates of vehicles were copied by a federal subcontractor and stored on a network that was subsequently hacked [42]. The 2019 data breach in the security platform BioStar 2 leaked 27.8 million records and 23 gigabytes of data including facial recognition data and user face photos. The platform is reportedly integrated into another access control system used by 5,700 organisations in 83 countries, including governments and banks [70].

the performance on the benchmark is likely to overestimate performance on the target population. This is because, in general, it is more difficult to distinguish among members of a more homogeneous population.

The difference in environmental factors between benchmarks and real-world applications is also critical. For instance, benchmark results for a face recognition system that uses a dataset of mugshots taken in controlled environments are not reflective of results obtained when deployed in an uncontrolled environment such as on a police body camera. The central problem is this. **There is a virtually unlimited set of conditions under which FRTs can be used, and standard, fixed benchmarks can only model a small number of these.**

Another issue is that gathering the volume and variety of data needed to evaluate a system robustly can be challenging when the process requires consent. Collecting images without consent poses privacy violations which comes with legal risks and ethical concerns. For example, one concern is predatory inclusion practices such as the harvesting of face data from vulnerable populations like homeless individuals [33].

A third issue concerns adaptation to benchmarks. That is, the longer a benchmark is publicly available, the easier it becomes for developers to produce systems that are uniquely tailored to the specific benchmark. Over time it is possible and likely that FRTs will be adjusted to maximize performance on a particular benchmark. This practice may show apparent improvements without making sizeable gains on previously unseen test data.

Some of these issues can lead to a false sense of progress and misleading interpretations of benchmark results. Additionally, despite the temptation to rely on a universal metric to make it easier to compare the performance of FRTs on different populations, no single number can fully convey how a system performs. There are different performance measurements for a system which focus not just on total accuracy, but also which types of errors it makes, how long it takes to produce a result, and how robust it is to varying conditions.

Finally, despite high accuracy numbers reported on standard benchmarks, there are no guarantees (currently) that FRTs will not produce errors due to other factors. When it comes to hardware, cameras can produce blurred or low-resolution images. The angle at which a face is captured can make it more difficult to extract facial features. For example, profile images where a subject is facing left or right generally provide less information than front-facing images. The amount of light available when capturing a face image influences how much information from the face can be adequately captured. Furthermore, accessories like hats, masks, and scarfs can block or cast shadows on key areas of a face making performance less reliable.

Currently, there is little guidance from manufacturers about the conditions required to achieve high accuracy of FRTs. Even more importantly, there is little to no discussion of how a user of a system can know when the system has been presented with an image whose quality is inadequate. Systems must be tested not just for their accuracy on high quality images, but for their ability to reject images that are below the quality needed. **That is, a safe and effective system should report that it is not able to make a decision due to the low quality of an image, rather than simply giving its best guess.** To date, this aspect of quality assurance is severely underdeveloped.

Lastly, when it comes to differentiating human faces, there are known cases where FRTs especially struggle. Identical twins provide a common example of two different faces that are very similar and are difficult to distinguish, both for humans and for machines. Tests also show that the faces of babies and children can be harder for machines to distinguish than their adult counterparts. As people age, hormonal changes, illnesses, and injuries can alter facial structure, making it more difficult to identify a face from older facial images.

These factors do not capture the full complexity of developing and measuring FRTs. They do reveal why we cannot simply assume that technologies have matured based on reported accuracy numbers and benchmark results. Poor performance on a benchmark can serve as a warning flag, but good performance on a benchmark is not a green light. Critically, we cannot rely on these numbers to determine the suitability of using FRTs for specific applications.

2.3 Legal challenges: Existing protections and legislative gaps

A variety of factors have motivated the passage of local and state laws addressing FRTs. These include the threat of mass surveillance, privacy violations, and disparate impacts on marginalized groups. At present, federal US laws regarding FRTs address different applications, have different jurisdictions, and cover different time periods. They address only a small subset of the potential uses of FRTs.

The ability of FRTs to develop profiles of individuals based on facial analysis also raises legal concerns pertaining to discrimination and privacy. We focus here on anti-discrimination law and constitutional rights that are particularly relevant to current and emerging applications of FRTs. We outline challenges presented by existing and proposed legislation for FRTs in the United States.

Anti-discrimination laws relevant to FRTs. Evaluations of permissible use of FRTs must address the risks of violating existing anti-discrimination laws. In particular, *Title VII of the Civil Rights Act of 1964* prohibits discrimination on the basis of race, color, religion, national origin, or sex. FRTs used in employment and housing domains risk class action federal lawsuits if they are found to perpetuate or mask discrimination on the basis of a protected characteristic.

In addition to sexism, racism, xenophobia, and religious persecution, federal law also offers protection from ageism in the workplace. The *Equal Pay Act of 1932* prohibits discrimination based on age. The ability of FRTs to assess age can be used for intentional age-based differentiation that can be litigated as discriminatory based on disparate treatment.

Another area of discrimination with particular relevance to facial recognition technologies deals with ableism. *Title I of the Americans With Disabilities Act of 1990* prohibits discrimination against a qualified individual with a disability. FRTs currently rely on training and test data that seldom include individuals with a range of disabilities providing little knowledge about the performance of people who are differently abled. A minimum requirement for the deployment of FRTs should include processes that check adherence to established anti-discrimnitation laws.

Constitutional rights relevant to FRTs. Constitutional concerns around FRTs cluster around civil rights established by the First, Fourth, and Fourteenth Amendments. The *First Amendment* provides freedom of religion, freedom of assembly, freedom of petition, freedom of speech, and

freedom of the press. FRTs, when used for surveillance, can track where people go to worship, the people they associate with, and whether they attend a protest. Knowledge of ongoing face surveillance can inhibit a person from exercising their First Amendment rights for fears of retaliation and social stigma [26].

The *Fourth Amendment* prohibits the search and seizure of a person or their artifacts without probable cause. Deploying a surveillance camera with FRT capabilities risks violation of Fourth Amendment protections that would require a warrant to search for an individual. Even if a warrant is obtained for one person, in order to try to find that person all faces detected in a camera feed may be algorithmically searched without consent.

The *Fourteenth Amendment* provides all US citizens equal protection under the law and the right to due process. FRTs that have been shown to have racial, gender, and age bias spread associated risks unevenly. Thus, when applied in areas like law enforcement they can be argued to be in violation of equal protection under the law. Due process requires having access to the evidence used to make decisions. When FRTs are used covertly, due process rights can be violated. For instance, this can occur when a system is used to inform an investigative lead but this fact is not revealed. When FRTs are incorporated into employment decisions like hiring, an individual has a right to know what technologies of consequence were used in order to push for redress of harms. Without disclosure of use, there cannot be due process. When applications of FRTs have credible risks to civil rights, such cases must be subject to scrutiny not only from individuals with technical expertise but also from representatives of the impacted communities.

Limitations of existing and proposed legislation for FRTs. The passage of city and statewide bans and moratoriums on the use of FRT, along with proposed federal legislation (e.g. the *Commercial Facial Recognition Privacy Act of 2019*, the *No Biometric Barriers to Housing Act of 2019*, and the *Facial Recognition Technology Warrant Act of 2019*), show urgent public concern about privacy, consent, discrimination, and surveillance. They also show the need for comprehensive policy to address the wide range of uses across private and public sectors. Legislation that is domain specific, regionally placed, and time limited leaves many applications and deployment areas unaddressed. All of these areas need oversight mechanisms and guidance for industry practice, research norms, and procurement procedures.

Domain specific laws that address exclusively either public or private sector use can leave unaddressed the critical interface with private companies and vendors supplying government agencies with FRTs. Private companies that operate internationally have no obligation to remain loyal to US interests. When government agencies use unregulated FRTs, the data they submit can be used in unknown ways by the companies providing the services. Laws that focus on federal levers to enact restrictions, like those proposed to put a moratorium on the use of FRTs in federal housing, can provide a buffer to unregulated and unwarranted face surveillance for some communities. However, the widespread use of surveillance systems with FRT capabilities in homes owned by private individuals also needs to be addressed.

Regionally focused legislation has been passed to govern FRTs, yet the lack of federal laws and the narrow focus of local laws leaves the vast majority of the country without guidance or protections. Across the United States, cities and states are enacting laws that put restrictions on FRTs. Washington became the first state to pass legislation outlining how and when FRTs can be used by law enforcement. In Illinois, the enactment of the *Artificial Intelligence Video Interview Act*

requires employers to disclose use of FRT in hiring, obtain consent, and engage in data minimization practices. Cities including San Francisco (CA), Oakland (CA), Cambridge (MA), Brookline (MA), and Springfield (MA) have approved moratoriums on government use of FRTs. At the same time, vendors of FRTs operate in a global landscape across private and public sectors. An FRT application developed overseas that is used for entertainment purposes can lead to data breaches that can be as concerning as data breaches of government data. The global landscape of these technologies requires thinking through not just how to mitigate risk at city, state, and local levels but also federally and internationally.

Finally, the temporary nature of many bans (often with time limits between one and five years) introduces further complexity. What happens when they expire? Does the expertise needed to evaluate them need to be reassembled each time these laws are reconsidered? Such temporary bans at different levels of government buy time to consider further implications of the technology, but do not represent a good long term solution. A federal office where these important issues can be considered together would be both more effective and more efficient than the current ad hoc responses. In total, the legal landscape, the technical challenges, and the societal challenges discussed here motivate the establishment of a federal office to manage these complexities.

SECTION 3

The FDA: A Precedent in the Regulation of Complex Technologies

In the previous section, we enumerated many of the common challenges associated with real facial recognition technologies (FRTs). We also discussed some of the previous efforts to address these challenges through new datasets, principles, and new laws, and why we believe these are not sufficient. What then, is the right way to manage this complex industry.

One path forward is to examine frameworks that have been established over many years to handle other complex technologies that address the tension between harms and benefits, advanced and changing technological issues, and complex legal and ethical concerns. There are many possible examples of such frameworks, but here, in particular, we consider some of the lessons that can be learned from the US Food and Drug Administration (FDA), an enormous organization developed over more than 100 years to perform the difficult job of regulating medications and medical devices in the United States.

Below we will review some of the structures that have been developed by the FDA and related agencies, and by related laws, policies, and conventions. We repeatedly ask the following question:

Question 3.1

Does a particular process, convention, law, policy, or regulatory structure that is used by the FDA suggest an analogous mechanism that could play a useful role in the regulation of facial recognition technologies?

We shall argue that the answer is often a resounding “yes”. In other cases, there may not be a direct analogy, but an underlying goal that is shared between the problems of medical regulation and the regulation of FRTs. We shall then consider what types of mechanisms may help us achieve similar goals to those issues that have been dealt with by this very large and successful organization. We make many specific recommendations about mechanisms that can be borrowed from these existing policies and procedures.

To give the reader a sense of where we are going, consider some of the following conventions, rules, and procedures that have been set up around the regulation of pharmaceuticals and medical devices in the US:

- Before a medical device can be marketed in the US, it is classified into one of three major categories according to the risks associated with its intended use. The same physical

device may be categorized differently depending upon its intended use.

- There are a variety of gatekeepers in the US medical regulatory system that help ensure that proper considerations are made at various levels in the regulatory process. These include:
    - Prescribers of drugs, who must be medical doctors, that determine the people who may be the recipients of certain drugs.
    - Regulators (at the FDA) who determine whether a device is ready to be marketed in the US, based upon data provided by manufacturers.
    - Pharmacists, who have tight control over the transferral of drugs from the manufacturers to the patients, and whose behavior is gated by the prescriptions from physicians.
    - Registered manufacturers, who have been given the legal right to manufacture certain drugs or medical devices.
- When a new medical device is developed that has no clear precedent in the marketplace, an elaborate process of defining and executing a study through an Investigational Device Exemption (IDE) is followed, in order to gather data about a new device in real scenarios under the careful watch of an Institutional Review Board (IRB) and the increased reporting requirements of such a study.
- When a medical device malfunctions, manufacturers, device user facilities, and importers are required to file reports to the FDA. The FDA also encourages health care professionals, patients and caregivers to submit voluntary reports about adverse effects.

These systems were not developed quickly, but rather are the result of a mature system which has responded to a series of problems over many decades, and were developed to address these problems, conflicts of interest, and to balance the forces at work [5]. In the remainder of this chapter, we take a closer look at some of these processes, and what they might have to say about a way forward for the complex world of FRTs.

3.1 The FDA's management of the medical industry: Key concepts

The US FDA is an enormous organization with a complex hierarchical structure. Rather than detailing this exact structure, we extract some principles relevant to the management and regulation of FRTs. Many of these are successes, but some are also failures. We discuss both successes and failures below.

3.1.1 Indications, counter-indications, and intended use

One idea for an approach to managing FRTs is to put in place rules that guarantee its performance in all possible situations. Such situations include, at least,

- variability of subjects on which it might be applied, including differences in age, gender, appearance, skin tone, hair styles, accessories, and so on;
- variability of the conditions under which it might be applied, such as lighting conditions (day vs. night) or weather (rain, snow, bright sunshine); and
- image quality, including low resolution versus high resolution, blurred images (for example due to motion of the subject or the photographer), black and white versus color sensors, infrared cameras, or distorted lenses.

However, experience shows that in any one of these categories, there are many situations in which it is unrealistic, at least for the foreseeable future, for FRTs to perform well. This suggests an alternative approach to regulate the conditions under which FRTs might be applied.

Considering an analogy with the FDA, imagine for a moment that the FDA would not approve a medication unless it could be taken by anyone at any time under any conditions. Even relatively benign drugs such as aspirin or ibuprofen frequently include

- warnings for those with stomach ulcers and other pre-existing medical conditions;
- prohibitions for those under a certain age or weight;
- limitations on the rate or duration for which the drug should be taken.

Rather than developing drugs that are safe and effective for anyone in any condition, a completely different approach is taken – clearly describe the "indications," i.e. the allowable uses of a drug, and the "counter-indications," the situations in which a drug should not be used.

It seems fitting to apply the same type of reasoning to FRTs. Software developed for one setting should not be used in another setting. Software only tested on one population should not be used on another very different population. These ideas lead to our first recommendation.

Recommendation 3.1

Require manufacturers to carefully specify the intended use of each separate application of FRTs. This intended use will be at the center of a variety of processes, including the categorization of the product, the level of risk it entails, the validity of a deployment, assessment of misuse, and more.

While in some cases, manufacturers make suggestions about how their software should and should not be used, there are many cases in which such recommendations are completely absent. In addition, it is critical not just to specify intended use, but to make sure that operators are aware of these, that they are trained, and that there are real consequences for violating them. One idea is to establish legislation that makes it illegal to use such applications in violation of the intended use. This is one way to allow the use of carefully managed software rather than banning all potential uses. In Section 4, we give a detailed definition of intended use in the context of FRTs.

3.1.2 Classification by risk

Once one has established the intended use of a medication or a device, the groundwork is laid for assessing the risk. For example, devices made to be implanted in humans, such as pacemakers or stents, represent far greater risk to patients than devices that are only temporarily in contact with a patient, such as bandages or blood-pressure monitoring devices. Common sense suggests that high risk devices warrant a more thorough and extensive review than low risk devices.

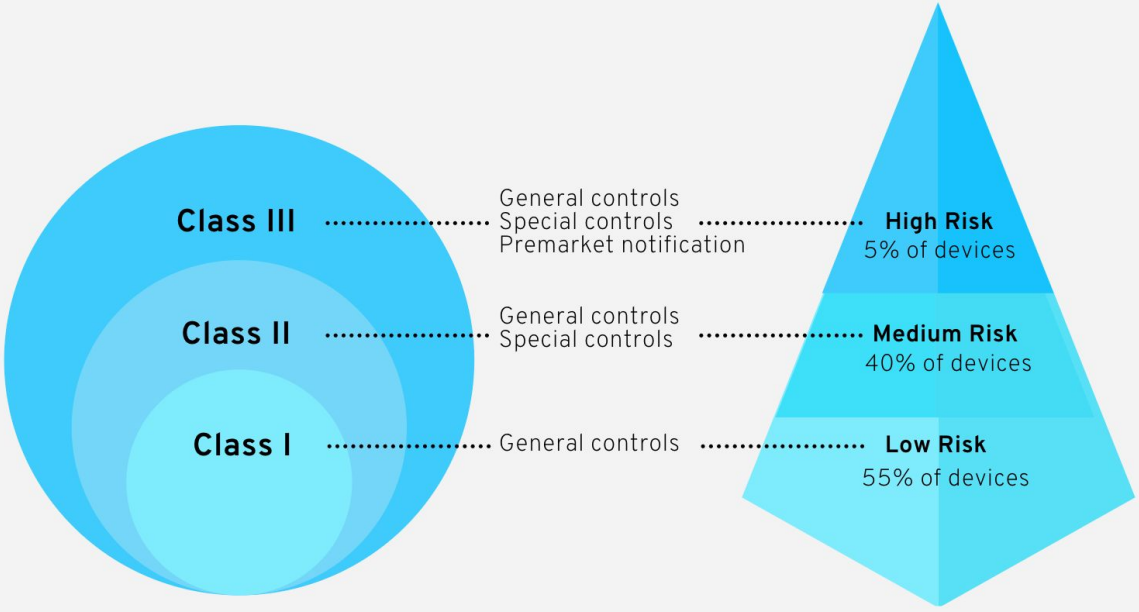

Figure 2: The hierarchical structure of medical device regulation at the FDA.

We recommend the same principle should be applied to FRTs. Surely it makes sense to subject an application intended for use in identifying criminal suspects to more scrutiny than a recreational phone application used for enhancing personal photo portraits. In these two applications, the costs of errors are extremely different, and suggest different levels of analysis. The FDA has

classified medical devices based upon their risk (informed by the intended use), and these classifications help simplify and improve their regulation.

In particular, the FDA has established three classes of medical devices based upon the inherent levels of risk. Simple objects like tongue depressors are classified as low-risk, Class I devices. On the other end of the spectrum are Class III devices like cardiac pacemakers, whose malfunction can quickly lead to serious injury or death.

A primary purpose of the separation of devices into risk categories is to facilitate the tailoring of regulation to the inherent risks. Simply put, riskier devices need more oversight and regulation. Figure 2 shows how each class of devices at the FDA inherits the risk management from the lower level, but adds a new level of regulation to account for the greater risk.

Recommendation 3.2

Organize controls for FRTs in a hierarchical fashion as a function of the risk level for each category.

In Section 4 we will offer an initial classification of FRTs into risk categories. However, this is just a starting point. Establishing the appropriate level of regulation for each class of FRTs will be an ongoing task that incorporates inputs and observations from manufacturers, users, subjects of facial recognition, the general public, civil rights advocates, lawyers, law enforcement, and many others.

3.1.3 Approval for commercialization

When new drugs are developed, they cannot be marketed or sold in the United States until they have been assessed by the FDA. This is a complex process that differs for over-the-counter and for prescription drugs. Often, a study specific to the drug at hand must be completed, analyzed, and reported to the FDA for analysis. One such process is called a "New Drug Application" or NDA. Companies pay fees to support such processing, which both help to support the required analysis and limit frivolous or high risk submissions which are likely to be rejected.

Thus, before the management of how drugs are used by individuals becomes relevant, there is a process which governs whether they can be used at all. In the pharmaceutical world, this is a complex and laborious process that can take years to complete. Clearly there are important trade-offs to balance: the possible benefit of having a new drug that could improve patient health with the risks of side effects, overdoses, and unintended use that might come from rushing things to market.

Managing the trade-off between benefits and risks has been a continual challenge for the FDA, and has led to new types of processing such as *Accelerated Approval* for urgently needed drugs (for example, a drug intended to help in an epidemic) and new fine-grained categorization of drugs based upon the particulars of the approval process. A second trade-off to consider in both regulation of drugs and in the regulation of FRTs is the balance between complexity of regulations and efficiency of implementation. For example, the FDA has in-house expertise in chemistry and medicine that lets them understand where the "danger points" of certain drugs are. Thus, they can adapt their requirements for testing to individual drugs, their chemical

formulas, and their likely side effects. Such adaptation requires funds to have the appropriate expertise on board, but results in major efficiencies so that unnecessary studies are not done, and resources can be dedicated to the most important questions specific to a particular drug.

Such expertise in evaluating facial recognition algorithms could similarly improve the efficiency with which certain pieces of software are approved, but may come with significant infrastructure costs. The FDA model in which companies who have an interest in rapid approvals can at least partially support the financial burden of the FDA's independent investigators is a way to fairly improve the efficiency of these types of organizations.

Recommendation 3.3

An organization to regulate and manage FRTs should be both independent of industry and have sufficient expertise in facial recognition to evaluate the safety and efficacy of carefully-specified applications.

3.1.4 Approval for specific use cases

When a drug is approved by the FDA for marketing, it does not mean that it can be used by anybody. An additional level of regulatory support is needed to determine in which specific cases a drug may be used.

Within the FDA system, medical doctors have a special role in that they are the gate-keepers for FDA-approved prescription drugs. As a group, they have been trained to evaluate whether certain medications are appropriate for certain patients with certain conditions. Indeed, many medications are typically only prescribed by those with additional special training. For example, drugs that modulate hormone levels may be prescribed primarily by endocrinologists who specialize in understanding and regulating hormone levels.

There is a great deal of infrastructure inherent in this gate-keeping. There are specific requirements about education, training, and certification that must be satisfied to gain the right to "practice medicine" in the United States, which allows the prescription of medications. Some medications have such low risk, and are safe for such a large segment of the population, that it has been deemed unnecessary for them to be controlled by the prescription approval process. This has led to the classification of over-the-counter medications, which allows the general public to benefit from lower risk medicines such as aspirin without the time and expense of a doctor's visit.

Finally, it is critical to note that the following are distinct processes: the clearance of drugs for marketing and sale in the US and the prescription of drugs to an individual patient. In the pharmaceutical world, one or the other of these is not enough–it is critical to have oversight at the research and development, production, packaging, and distribution levels (FDA clearance), and also at the level of distribution to the individual patient. These two levels of oversight are handled by completely different groups of people, with different tools, skills, and degrees of effort.

Recommendation 3.4

For high risk applications, establish an additional level of oversight for FRTs to oversee individual deployments. These ***individual deployments*** *are specific distributions of software in a new setting, analogous to the prescription of an approved drug to a new individual.*

Consider the case of applying FRTs for voluntary screening of a medical condition (See Section 4.4 for a detailed example). Even after such a system has been approved generically, there are special circumstances that affect each individual deployment. Critical questions that affect the appropriateness of each deployment include (but are not limited to) the following.

- Does the software and the procedures for using it comply with relevant local and state laws?
- Do the instructions for the system support the most common local languages, providing access to non-English speakers?
- Are the communications systems over which diagnostic data are sent secure to ensure the privacy of individuals?

We argue that systems such as this must be re-evaluated for each individual deployment in order to insure its safety and efficacy in each environment. The degree of analysis required for individual deployments should be tailored to the system's risk category.

3.1.5 Adverse effects reporting

When a patient experiences an adverse effect from a medication or a medical device malfunctions, there are parties (such as hospitals) that are required to report these problems. In addition, the FDA welcomes voluntary reports from others, such as patients that experience adverse effects.

Once again, it seems reasonable, especially for applications in which errors have serious negative consequences, to require such reporting for FRTs. This leads to the next recommendation.

Recommendation 3.5

For medium risk and high risk deployments, manufacturers and users should be required by law to keep detailed statistics on any known false positive identifications, false negative identifications, reported harms, and other adverse effects.

At a minimum, such requirements help manufacturers learn about potential problems with their systems and rectify them for future releases. In cases where persistent patterns of erroneous

behavior emerge, it may be appropriate to issue recalls until the problems are addressed.

3.2 A chance to do even better

The FDA and the medical community as a whole have done a remarkable job in improving the production, distribution, and management of pharmaceuticals and medical devices in the US, and similar systems have been adopted in many other countries and in the European Union. These systems have been so effective that many people simply take it for granted that medical devices and drugs are safe and effective.

However, one aspect of medical research has been identified as problematic: the limitation of the study of medicines to narrow cohorts. Historically, many phases of the drug approval process explicitly required the exclusion of many groups, including young patients, women of childbearing age, and others. In particular, many studies of pharmaceuticals have been restricted to populations that are not representative of the ultimate target populations for a drug. For example, recent studies have found that drugs for treating asthma in children have been primarily developed by studying effects on Caucasians and may be poorly suited to treating African Americans [38]. Recent research shows that pharmaceutical studies need to better reflect patient populations to understand differences in how various populations respond to various medications [53]. This is clearly an issue both of public safety and fairness.

3.2.1 Fairness in drug regulation and facial recognition technologies

Similar issues have been identified early on as significant problems in the deployment of FRTs. The variable performance of systems on different sub-populations is now well-documented. In 2012, Klare et al. [44] performed an extensive study of the role of demographic information such as age, gender, and race in the matching accuracy of facial recognition methods. This study showed that three commercially available systems had lowest accuracy on faces of people identified as Black for their racial group.

A later report released in 2015 by the National Institute of Standards and Technology (NIST) [52] showed that gender classification from faces consistently showed higher accuracy rates for male faces than for female faces, for a wide array of systems submitted for evaluation. A more recent study by Buolamwini and Gebru in 2018 [16] compared gender classification accuracy rates for publicly available systems across gender and skin types, and showed that dark skinned women were recognized as female at consistently lower rates. The December 2019 NIST Face Recognition Vendor Test on demographic effects in face recognition systems showed "the majority of face recognition algorithms exhibit demographic differentials. A differential means that an algorithm's ability to match two images of the same person varies from one demographic group to another" [31]. More broadly, issues of bias have been identified in a range of related artificial intelligence technologies, including decision making [77, 64], interpreting natural language automatically [15, 17], machine translation [68], and machine vision [78, 35, 74]. These findings are similar to the findings in FRTs: these technologies tend to encode biases with respect to protected demographic categories.

Addressing such fairness issues is one place where we have a chance to do better than the medical industry from the beginning. In particular, the differential impact of FRTs on different populations should be considered both from a usage point of view (see Section 2 for more), and from a fundamental technology perspective. That is, FRTs may impact different groups in different ways for two reasons:

1. Because the technology exhibits fundamentally different accuracy on different sub-populations (analogous to differential effects of drugs on different groups).

2. Because of bias in how it is used or interpreted by users of the technology.

This leads to our final recommendation of this section.

Recommendation 3.6

Include in assessments of FRTs both an analysis of the underlying technology, and an assessment of the risks for different populations, accounting for the circumstances in which the systems are deployed and used.

3.3 Summary

The recommendations in this section are a starting point for the discussion of concrete steps that can be taken to manage the complex ecosystem of FRTs. Such large-scale infrastructures to regulate complex technologies do not emerge overnight—the FDA has evolved continuously over many decades. Similarly, we expect to incorporate feedback and discussions about these ideas in order to improve them and adapt them to the regulation of FRTs.

SECTION 4

The Structure of Facial Recognition Regulation: Core Definitions

In the previous section, we introduced a number of key concepts established by the FDA in its management of pharmaceuticals and medical devices. In this section, we more fully develop analogous concepts for the management of facial recognition technologies (FRTs).

In particular, our goal is to provide a starting point for the management of a large family of technologies around facial recognition. This requires careful definitions, conventions, and processes. In this section, we give an overview of our approach, and provide some definitions of key concepts that we wish to make precise. These clearly-defined concepts will provide the foundation necessary for beginning to talk about such a system of management. The goal is not to completely specify how such a system would work, but to illustrate how certain issues can be addressed through processes similar to those in other regulatory frameworks.

4.1 Overview

FRTs represent too many applications for a single set of rules. Just as different restrictions are applied to different medications, FRT controls should be tailored to the application. This requires mechanisms for carefully defining the scope of applications. The terminology developed below will be used to define the scope of FRTs, and will subsequently be used to assess risks and manage these applications in real-world settings.

As discussed previously, the concept of intended use is central to our framework. A rigorous and detailed definition of **intended use** (Section 4.2) will, among other aspects, describe how, where, and for what purpose a facial recognition system is deployed. Because the same facial recognition system can be deployed many times, deployment is a multi-level concept. There is the question of what **type** of deployments are intended. In addition, there is the question of whether an **individual deployment** is reasonable.

As an example, a system might be certified for "deployment in retail stores". We consider this to be the deployment type. It describes the general type of setting in which a particular system is deployed, and general parameters of use. However, even after being certified for such a deployment type, it is important to assess an **individual deployment**, such as the deployment to a *specific retail store*, for appropriateness.

The FDA's drug approval process and the prescription of an approved drug provides a direct analogy.

- The approval of a drug by the FDA for specific indications is analogous to the approval of a facial recognition system at the level of the **deployment type**.

- The prescription of an approved drug by a physician for a particular individual and for the treatment of a specific condition is analogous to the approval of a system, approved at the level of **deployment type**, for a specific **individual deployment** (e.g., a specific retail store).

Once an intended use has been carefully defined, we can ask whether a piece of software is being used in accordance with its intended use (a **valid deployment**) or in violation of its intended use (an **invalid deployment**). In the medical world, this is analogous to the terms on-label use or off-label use which indicate whether or not a drug is used in accordance with its intended use.

Given this infrastructure, the next step is to consider the following two critical questions:

- If a face recognition system is a **valid deployment**, what risks are associated with its use? This question should be carefully explored by the manufacturer, with supporting data and experiments that are evaluated by regulatory experts.

- In addition, we must ask, what is the risk that such a system will be used in an **invalid deployment**, and how might that occur? Thus, we must explore not only the risks of a system when used as intended, but the precautions necessary to minimize the inappropriate use. Such analyses are seen, for example, in the analysis of new pharmaceuticals for the potential of abuse.

In the following sections, we revisit the fundamental terms of *intended use*, *deployment type*, and *individual deployment*. Our goal is to expand and clarify their meaning in the context of FRTs and to provide a set of working examples which make it clear how these terms can allow us to manage the risks and benefits of these technologies. We also discuss how these definitions allow us to categorize different types of applications, assess their risks, and lead to appropriate controls for each category. We start with the central concept of intended use.

4.2 Intended use and its specification

Clearly defining the intended use of software is essential in understanding the implications of its use, exploring possible risks, and implementing appropriate controls that are commensurate with the level of risk. No facial recognition application can work in all settings. While each application is different, they all have limitations. Examples include the following.

- Applications will require some minimum degree of image contrast (a measure of the difference between the brighter and darker portions of an image) to make accurate identifications. Some applications may require more contrast than others or have higher accuracy requirements than others. But for all applications, error rates will rise dramatically as image quality drops below some critical level. This level will depend upon the application.

- Applications may have been developed to recognize adults or people of a certain age, and fail to recognize children at a sufficiently high accuracy. Other applications may be explicitly designed to have high accuracy on children. Application error rates may go up dramatically when applied to a subgroup for which the application was not designed.

- An application may have been designed for a setting in which errors have only minor consequences, such as sorting personal photo collections by identity. If such an application is used to make decisions in which errors have more severe consequences, such as matching a person to a felony database, the risk increases dramatically. Thus, it is critical that applications only be used for the decision support that was originally intended and planned for.

These are just three examples of the relationship between the settings in which software is used and the safety and efficacy of a particular FRT application.

Goals of the intended use specification. The specification of intended use serves several purposes. A well-designed FRT system not only has a clear specification of the settings in which it is designed to perform safely and effectively, but also considers how to handle situations in which an application is misused, in violation of the intended use. For example, no system can be expected to recognize a person from a completely blank photo and this should be made clear by the intended use. But, in addition, a well-designed system might let the operator know if an input was blank and outside of the intended use rather than simply returning an incorrect identification. That is, a system should identify, either automatically or by human inspection, when the image it has been given is outside of the range of intended use. Today, many FRTs make assumptions that the images they are working with are of adequate quality, and do not perform such checks.

The specific goals of the intended use specification include the following.

- To specify a regime of parameter settings and conditions in which the software has been shown to work safely and effectively. This corresponds to the innermost region in Figure 3. The regime of safe and effective parameter settings and conditions should be reflected by product warnings, instructions, and in operator training.

- To provide requirements for a process of *intended use verification*. This process, which may be fully automatic, partially automatic, or manual (i.e., performed by a person), is intended to verify, for each use of the system, that it is being used in accordance with the specified intended use. Alternatively, one might give assurances that certain unintended uses pose negligible risk.

- To provide context for risk analysis. The risk analysis should include
    - An assessment of the risks given that the application is used as intended.
    - An assessment of the probability that intended use will be violated.
    - An assessment of the risks given that the application is used in violation of the intended use.

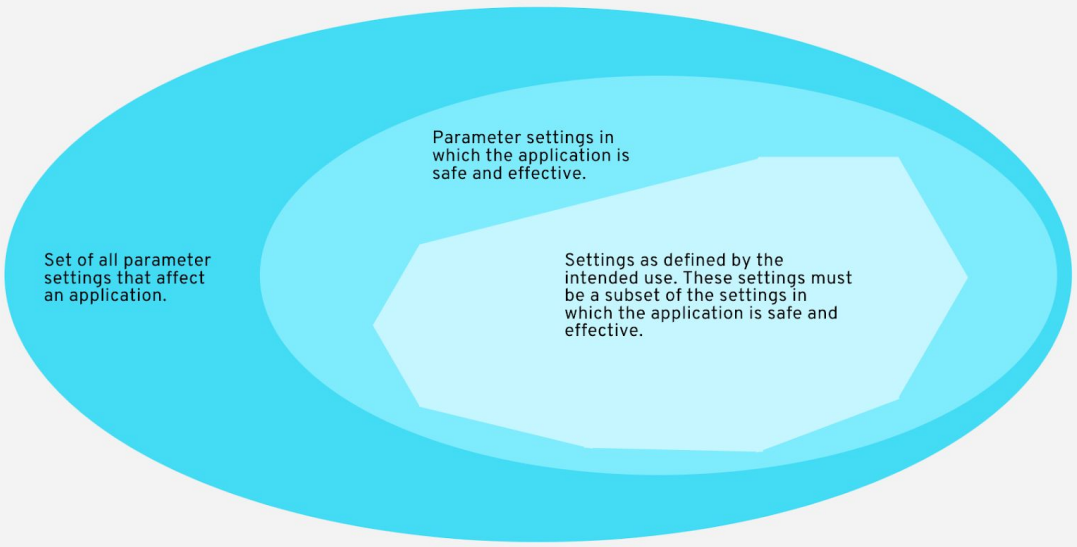

Figure 3: The specification of intended use.

This figure is an abstract representation of all conditions (or parameter settings) under which an application could be used. Only for some of these settings (the intermediate region) is the application safe and effective. In settings that are outside this region, the application should not be applied. The intended use (innermost region) should be defined so that the intended parameter settings are contained strictly within the region of safe and effective use. There will be a natural trade-off between the ease of defining such a region and the degree to which it allows the largest number of uses. That is, one might sacrifice certain possible uses of a system in order to make the intended use simpler and easier to describe and manage.

The intended use of a system and its risk analysis should be developed in tandem. An initial description of intended use will influence the risk analysis and may suggest new restrictions to the intended use. After modifying the intended use to reflect these risks, the risk analysis should be updated. This cycle should be continued until a well-defined intended use and a paired risk analysis are consistent.

While this white paper is not intended as a complete guide to building safe and effective FRT systems, we believe it is worth developing concrete elements of an intended use definition, and start with the definition below.

Definition 4.1

*The **intended use** should specify, as precisely as possible, the conditions under which a system could be used and also the decisions which may be supported using the system. The details of such a specification may be related to the general risks of the application, but should consider, at a minimum*

- ***Who** the intended target population is. These are the people that the system is intended to analyze or recognize.*
- ***Who** the intended users are. These are the people who will use the system, and whose decisions will be affected by the system.*
- ***Where** the system can be used. This includes geographical locations and legally defined regions with potentially different laws.*
- *The **conditions of use**. This includes factors such as weather, lighting, temperature, and distance from subjects. These factors are distinct from the image quality requirements below, and should be considered separately.*
- ***Image quality requirements.** This includes topics related to image formation such as inherent resolution, distance to subject, availability of surrounding context, motion blur, camera specifications (such as lens quality and sensor properties), and time of exposure.*
- ***Decision support.** This requirement specifies the intent for supporting decision making. For example, a system may be prohibited from identifying a person, but only be used to count people (for example, in crowd control applications).*
- *Specification of **deployment type**. Does this application represent an existing deployment type or a new deployment type? (See Section 4.3.)*
- *Specification of **individual deployments.***
- *Specification of **counter-indications,** i.e., explicitly prohibited uses. This allows manufactures the opportunity to detail ways the technology should not be used.*

In Section 4.4, we give an example of a simplified intended use definition for a real-world example. We now move on to discuss deployment types and individual deployments.

4.3 Deployment types and individual deployments

In this section, we discuss the rationale and definition of **deployment types**. Deployment types are meant to represent general categories of FRT applications. As discussed below, such categories can be used to make reasoning about and management of FRT applications more efficient and more effective. Typical examples of deployment types might be:

- systems deployed within an office building for monitoring the people within the building;
- software installed on a personal computer for organizing personal photo collections;
- systems deployed in a federal forensic laboratory for identifying crime suspects.

These deployment types are meant to allow the grouping of similar FRT applications into broad categories for improved efficiency and targeted analysis.

Definition 4.2

*The **deployment type** is a grouping of FRT applications that have similar intended use and function.*

Possible benefits of such groupings include the following. When a new FRT application is analyzed by a regulatory office, previously analyzed products of the same deployment type (both approved and rejected) could provide precedents and models for such an analysis, allowing the new product to be analyzed more efficiently. Expertise for regulation, for purchasing, and for consumer groups, can be organized according to deployment types, making it easier for a regulatory office to organize and acquire the required expertise. Unexpected problems or consequences found after the deployment of FRT systems could be more easily applied to the analysis of subsequent applications if appropriate groupings are made.

Deployment types and risk categories

Another primary role of deployment types is to aid the categorization of FRT applications by degree of risk. While each FRT application is different, we assert that it is useful to form broad categories of applications according to their risks. For example, consumer applications that are only used for entertainment purposes are, if used in accordance with the intended use, likely to pose fewer risks than applications used in law enforcement. While these risks depend upon the exact details of each application, these broad categories can serve to organize the extent of controls applied to various FRT applications.

We propose the following risk categories of deployment types and provide two examples of each:

- **Low Risk**
  - Applications for sorting personal photo collections.
  - Applications for adding digital accessories to face photos for personal use.
- **Medium Risk:**
  - Face verification for driver's license renewal.
  - Face verification to unlock consumer phones.
- **High Risk:**
  - Face identification for matching a suspect against a felony database.
  - Facial analysis software to analyze a subject's suitability for hiring.

Further work is needed to develop mechanisms for categorizing risks. At a minimum the analysis of risk factors per deployment should include consideration for societal harms like threats to civil rights, individual harms like privacy invasion, legal violations, and expected technical failure modes. See Section 2 for an overview or societal, technical, and legal challenges posed by FRTs.

We foresee a regulatory system in which one of the first steps taken by the manufacturer of a FRT application is to establish its deployment type which informs risk category. This could initially be suggested by the manufacturer, but would be done in consultation with a regulatory office. There are cases in which the deployment type might not be clear or may fit in multiple categories, and the ultimate authority for categorizing an application would rest with the regulatory office. This procedure mimics the classification of medical devices into risk categories by the FDA.

Individual deployments

As mentioned in this section's overview, a given FRT application would be regulated at two different levels: with respect to its general deployment type and with respect to individual deployments, defined here.

Definition 4.3

*The **individual deployment** represents a single instance of an FRT system running for a specific purpose.*

As discussed in Section 3, assessment of individual deployments concern the specifics of a new environment in which an FRT application is to be deployed. Before a system is deployed in a particular scenario, questions about the specifics of a given environment must be answered. This, again, is directly analogous to the medical industry's requirement that a physician oversee the prescription of a drug to a particular patient, even though the drug has been approved for general use. The relationship among risk categories, deployment types, and individual deployments is illustrated in Figure 4.

Figure 4: The relationship among risk categories, deployment types, and individual deployments.

The figure shows the three risk categories. Each deployment type is contained within a single risk category, and each individual deployment is contained within a single deployment type.

4.3.1 New types of deployments

The goal of this document is not to develop a complete process for the review and management of all FRTs. However, a few more details about how we foresee these terms being used is warranted. When a new application is presented by a company, the question immediately arises about whether this is a fundamentally new application, or whether it is of the same deployment type as a previously approved application. If the latter, then the evaluation of the previous product can serve as a starting point for the analysis. Methods used for the mitigation of risks in related FRT applications can serve as a model for new applications with the same deployment type.

However, when a fundamentally new application emerges, it may require additional efforts to ensure safety and efficacy. Such new applications may require various types of trials to demonstrate that risks are sufficiently addressed.

In the medical device industry, pre-existing devices with similar functionality and intended uses to a newly proposed device are known as *predicate devices*. The FDA relies heavily on manufacturers' comparisons of new devices to predicate devices to improve the efficiency and relevancy of their analyses. We believe similar efficiencies could be realized in the regulation of FRT systems by their categorization into deployment types.

4.3.2 Software libraries and related issues

Another question that arises in the regulation of FRTs is how to handle components of systems that are not themselves final products. For example, a manufacturer may produce only a small portion of the software (a software library) that is used in part of a larger FRT system. Our primary intent is to address the regulation of systems sold to and deployed by end users, rather than the components that go into them. However, we believe that the "ecosystem" of FRTs will naturally incentivize third party component manufacturers (whether these be hardware or software components) to produce safe, effective, and well-understood components.

For example, suppose that Company A is producing an FRT system that uses a software *library* produced by Company B. While Company B's product would not be directly regulated by our proposed office, any company that sold a product using Company B's components would have to provide extensive testing, analysis, and assurances that involved Company B's component. Thus, in order to provide a compelling product for integration into a larger system, Company B would be incentivized to produce detailed specifications and analysis of their software libraries to facilitate approval for the integrator's products.

If discrepancies or inaccuracies in such specifications led to the rejection of downstream products by the regulatory office, this would presumably damage the reputation of the component provider. Such interdependent relationships arise in other regulated industries such as pharmaceuticals and medical devices.

The fact that high risk applications require substantially greater scrutiny would also lead to third party providers in a range of risk categories. Low risk applications might be supported by software libraries that were tested on generic benchmarks and standard datasets, while high risk applications would require more thorough vetting, and would presumably command a higher price. Such tiers in the software industry allow entry points at different levels of risk, expertise, and start-up costs.

4.4 A detailed example: Automatic screening for a medical condition

To illustrate these ideas in action, we present a specific deployment type to highlight risks, benefits, and trade-offs that occur in real applications. One possible use of FRTs is to provide automatic screening for certain medical conditions. This is possible for conditions which affect the appearance of a subject's face. Such applications have been explored in both academia [45, 50] and industry [32]. Here,we consider the use of FRTs in the diagnosis of *hyperthyroidism*, a condition in which too much thyroid hormone is produced. Hyperthyroidism can produce changes in the appearance of the face that can be detected automatically, presenting the possibility of more efficient diagnosis of the condition. Figure 5 illustrates some of the visual symptoms that may occur in a person with hyperthyroidism.

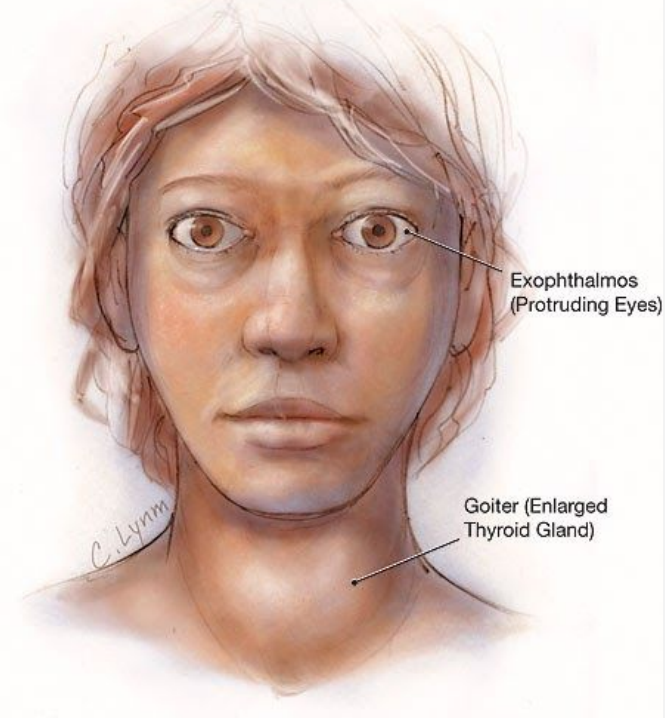

Figure 5: An Illustration of visible symptoms that may occur with hyperthyroidism.

Hyperthyroidism presents an example where visible symptoms of a disease may be recognized using FRT. Any claims that a system can reliably detect a disease by analyzing facial characteristics needs to be supported by comprehensive scientific evidence and empirical results showing effectiveness across a wide range of populations. The illustration depicted is reproduced from [71].

To make the example as concrete as possible, we consider the following specific scenario:

1. **Data collection.** During development of such a screening tool, images would be gathered of subjects both with and without hyperthyroidism.

2. **Application development**. Using these examples, a computer application would be developed to classify each individual as either positive (having hyperthyroidism) or negative (not having hyperthyroidism).

3. **Deployment.** After development, the application could be deployed for use. We consider a specific scenario in which free-to-use kiosks would be deployed in pharmacies for use by the public.

4. **Use scenario.** When a person approached the kiosk, they would be asked (with an on-screen message) whether they would like a free screening for certain medical conditions. Before starting, they would be informed about how the results of such a system would be processed (see below). If the person chose to opt in to the free screening, the system would take a photo of the person, and their photo would be screened for hyperthyroidism. If the application determined the risk of hyperthyroidism to be high, a private email would be sent to the primary care physician for further analysis. If follow-up was deemed necessary, the person would be contacted for a follow-up appointment with their primary care physician.

While there are possible benefits for this kind of application, there are also a variety of pitfalls and risks associated with deploying such a system. Our hope is that through a careful definition of intended use, a classification of the deployment type, an assessment of the risk level, and a

detailed analysis of the risks of a specific application, that we have a good start on how such a system should be managed and regulated. Below, we give preliminary examples of how the intended use, risks, and deployment types would be specified in our envisioned system. These definitions would be provided by manufacturers and presented to the office in support of a case that a system was safe and effective.

4.4.1 Automatic screening system: Intended use

Intended use component	Description	Mechanism to ensure
Intended subjects	Adults 18 or older. Must be able to either read or understand recorded instructions in deployment-approved languages. Must be pre-registered with primary care physician to obtain unique access code. Users must be able to press yes/no buttons or respond verbally to answer questions. System should accommodate subjects confined to wheelchairs. Subject must accommodate people with either visual impairments or hearing impairments, but not both.	Pre-registration with physician. Unique user number is assigned. User answers simple questions to confirm understanding of instructions.
Intended recipients of reports	Primary care physicians. Physicians must be certified in operation of the system, to understand the conditions under which a report is generated, their obligation to follow-up with patients, etc.	Physicians are pre-registered in system.
Where software can be used	Approved kiosks in approved locations. Must be available to the public.	Requires installation, inspection, and certification by manufacturer.
Conditions of use	Indoors in approved kiosks . Kiosk must have approved lighting (camera flash), neutral background, adjustable height mechanisms, headphones for the visually impaired, and easily accessible response buttons.	Manufacture approves location for use according to detailed environmental requirements.
Image quality	Kiosks must use cameras with approved resolution, exposure time, focal length and sensitivity.	The manufacturer may only use the system with a set of pre-approved cameras.
Decision support	The software is intended only for screening hyperthyroidism, Graves' disease, and related conditions. It is intended to detect visual symptoms of these conditions such as proptosis, goiter, and other visual symptoms associated with these conditions. When a patient is screened, appropriate outcomes are: 1) Above screening threshold, 2) Below screening threshold, 3) Procedure cancelled by user, 4) Procedure automatically cancelled due to other factors (power outage, inadequate lighting, etc.) If result is "above screening threshold", a report is sent via private electronic message to the primary care physician. The report includes the photo obtained by the system to make the decision.	The result of the system is only communicated to a pre-approved physician.
Counter-indications	The system is not intended to screen for medical conditions other than those listed above, even if visual symptoms are present. The system may not be used for any other purposes, including entertainment, photography, or diagnosis of unapproved conditions.	Such unapproved uses are minimized by the requirement for pre-registration with a physician, and the system's inability to print out photos or send digital photos to unapproved locations. Adequate labeling will ensure subjects understand the intended use.

Figure 6: Intended use specification.

The table gives an example of an intended use specification for an FRT system that performs screening for hyperthyroidism and related conditions.

Following the FDA model, we see manufacturers as the logical entities for defining the intended use. Intended use should be specified early in the design process, and modified if necessary as realities of development or risk assessment requires. A careful definition of intended use should

pave the way for risk analyses, testing procedures, appropriate documentation and training, and overall quality control.

By the time development is completed, the intended use should be well-specified and well understood by all parties involved from developers and engineers to salespeople and marketers. Proper labeling and training should ensure that those tasked with overseeing the systems and the users of the system are well-versed in what to expect. Figure 6 gives a preliminary example of the type of information that should be included in an intended use specification. In a real application, it would likely be substantially more extensive and detailed, with supporting documentation.

4.4.2 Deployment type and risk level

As described above, establishing a deployment type means defining a group of applications with similar goals and intended uses. For this case, we suggest a deployment type category of **public facial screening system**. Initially, such a deployment type might include any such system used for the public screening of non-communicable diseases (i.e., diseases that are not directly transmitted person-to-person).

Analysis of risks.

We envision that each deployment type would be classified according to a level of risk: low, medium or high. In order to establish a risk class, it is necessary to analyze risks inherent in such a deployment type. If it is later found that some deployments in this category represent high risks, and others represent low risks, it would be appropriate to create two separate deployment types, one at each risk level.

In analyzing the risks of this deployment type, a non-exhaustive list of factors to consider would include the following.

- Consequences of classification errors.
    - False positives. Risks of false positives include at least the following possible consequences:
        - ☐ Anxiety, stress, and stigma from a false diagnosis.
        - ☐ Costs to the patient, the physician, and others to follow-up with an unnecessary appointment. These costs may be both monetary and time-related.
        - ☐ Misdiagnosis by the follow-up physician. A physician may be biased towards a positive diagnosis by an incorrect result (a false positive) from the system. Such a misdiagnosis could lead to a variety of inappropriate, expensive, and potentially dangerous treatments.
    - False negatives. Subject fails to receive treatment for an existing condition.

- Consequences of different accuracy rates for different sub-populations.
    - Do lower accuracy rates for some groups shift the risk category to a higher level for this group?

    - Do lower accuracy rates for some groups create a significant reduction in value to this group, potentially rendering the service useless or even harmful?

- Process cancelled due to failure to capture an adequate photo that is automatically detected by the software. This might be due to inadequate lighting, subject motion, subject pose, subject position, or subject accessories (such as sunglasses).

- Other system failures.

    - Failure to communicate results to physician. This could result from incorrect information, out-of-date information, network failure, and many other causes.
    - Failure of physician to follow-up with subject in the case of a positive outcome.

- Misunderstanding of instructions or labeling by subject that can result from ineffective interface design or poor communication.

    - Misunderstanding of protocol, including what happens in the case of a positive or negative result.
    - Misunderstanding of privacy assurances.
    - Misunderstanding of operation instructions.
    - Misunderstanding of the reliability of the system. A patient may incorrectly interpret a negative result as a "clean bill of health."

- Misuse of the system for other purposes.

    - Recreation (children use the system to snap pictures of themselves, for example).
    - Medical data is stolen and used for other purposes, such as marketing or blackmail.

To simplify this example, we assume that the deployment type of public facial screening systems are categorized as having a medium level of risk . While false positives and false negatives could both be associated with substantial negative outcomes, such outcomes would be reasonable compared to an alternative where a subject has no access to free screening. For example, while a false negative result represents a case in which a patient with a medical condition is not diagnosed, they may not be diagnosed without the system in place to begin with. Of course, to claim that the benefits of such a system outweigh the risks, it would be essential to provide comprehensive scientific evidence and extensive empirical results showing the effectiveness across a wide range of populations.

It would be important to provide clear labeling and instructions to let subjects know that such a screening system in no way guarantees them a clean bill of health, and should not be used as a substitute for regular doctor check-ups.

The classification of a deployment type into a category of low, medium, or high risk would require extensive analysis and discussion which we will not engage in here.

4.4.3 Individual deployments

The general idea behind individual deployments is to focus on the specific needs of particular instances of an FRT system that are distinct from general considerations. For example, one deployment location may have a significantly different demographic than another deployment. Depending upon the risks of a given application, it may be essential to analyze how these

differences in demographics affect the safety and reliability of the system. As mentioned earlier in this document, we do this in analogy with the idea from the pharmaceutical industry that drugs must be approved for general use, but carefully prescribed by doctors for individuals.

The topic of individual deployments is complicated and a thorough discussion of it goes beyond the scope of this document. One reason for its complexity is that there may be FRT applications for which it is not necessary to define individual deployments and others for which multiple levels of deployment must be considered. Consider the following three examples.

1. **Sorting personal photo collections** (low risk). In this case, the harms due to errors *may be* sufficiently low that it is not necessary to adapt the analysis to separate deployments. In such a case, no approval of individual deployments may be necessary. This is analogous to over-the-counter medications, which require no prescription by a doctor.

2. **Medical screening kiosks** (medium risk). In this case, suppose that the performance of the system was deemed to depend significantly upon the demographics or predominant languages of the local population. In such a case, it would be appropriate to certify an individual deployment for each region deemed to have a distinct demographic make-up. This could be done at the level of cities, counties, states, or regions depending upon the specific analysis of risks.

3. **Deploying police body-cams for face recognition** (high risk). Such a high risk deployment includes not only the concerns about the demographics of subjects and abuses of the criminal justice system, but also the qualifications of each user (in this case, an individual police officer). It may be appropriate in such a case to require the certification of each individual user. In particular, it would seem appropriate to require each user to complete a training program in the risks and pitfalls of such a system. Such a system might require multiple levels of approval, from general product approval, to approval within a specific region by impacted communities, and final certification of individual users.

A key point in the analysis of such individual deployments is that the risk categories help streamline the process. It is appropriate to demand multiple levels of analysis for a high risk system that depends upon many local factors, and may vary for each user. At the same time, it is appropriate to bypass these procedures when risks are carefully analyzed and deemed to be minimal.

We believe that a single regulatory office where expertise, methodology, and institutional memory can reside is essential for implementing these ideas. Attempting to encode these into generic laws, relying on industry to do this on its own, or relying on local legislation alone to achieve these aims seems overly optimistic at best.

4.5 Summary

In this section, we have given definitions of key terms including intended use, deployment type, and individual deployments. We have also provided examples of how these definitions provide guidance to manufacturers in defining the scope of their systems, analyzing its risks, and providing mitigation strategies.

In practice, we believe that manufacturers should provide analyses of their own products, with definitions of intended use, categorization into a deployment type (or the proposal for a new

deployment type), and subsequent risk analysis. In addition, to be cleared for marketing, a manufacturer would naturally provide information about tests and their results.

While we do not delve deeply into the kind of testing we would expect here, it would depend strongly on the intended use and the risk category. For low risk applications, generic public testing, such as the NIST's Face Recognition and Vendor Tests may be sufficient. For higher risk applications, evidence that software is reliable in real-world settings, and the actual conditions of use would be expected. The nature and requirements for such testing are beyond the scope of this document. The key point is that these would depend heavily upon the application.

In keeping with the FDA model for drug and medical device regulation, we expect the manufacturer to present a coherent and well-argued case that their product is safe and effective for the given intended use. A panel of experts in FRT, its technical details and its pitfalls and societal risks, would analyze such a submission and grant market approval or deny it. Again, the specific procedures followed for such an analysis are beyond the scope of this white paper, but an understanding of these parallel processes at the FDA suggests many ways forward.

SECTION 5

Conclusion

We conclude by reflecting upon the FDA-inspired approach to regulate facial recognition technologies (FRTs). At a high level, we assert that the growing reach and complexity of FRTs necessitates 1) comprehensive oversight that can be provided by an office with federal authority and 2) dedicated expertise, not only of the underlying technologies, but also of the risks they pose in a range of application domains.

We have chosen to focus specifically on FRTs and not artificial intelligence more broadly. This focus allows us to provide specific examples, pitfalls, and guiding principles that are sufficiently detailed to inform practitioners and guide regulation. Much of our work here aims to illuminate the scope of FRTs; the task of precisely defining automated decision-making or AI-assisted decision-making is even more complex. If the focus is overly narrow, resulting recommendations may fail to apply to important domains on which these recommendations could shed light. Conversely, a diffuse topic reduces our ability to highlight specific technical challenges from a given domain. Our focus on FRTs allows us to provide concrete recommendations about documentation, categorization, and evaluation. We hope that the parallels among issues raised here and those that arise in related areas will serve as a template for efforts in other domains.

There would be many challenges in implementing our recommendations. To begin, while there have been visible examples of FRTs' shortcomings highlighted in both popular and scientific venues, much of the public is still unaware of the growing pervasiveness of FRTs. Like with the FDA's regulation of the medical industries, there will always be a tension between considering those with financial incentives for producing the technologies and those who benefit from oversight and regulation. Segments of the "tech" industry have opposed regulation within the US and abroad, citing concerns about hindering growth, innovation, and beneficial applications. Nevertheless, there is growing pressure for the regulation of technology companies, and even some evidence that the companies themselves would like some regulatory guidance, if for no other reason than to ensure a level playing field among the firms [72].

And so, despite the difficulties of implementing regulatory structure generally, and specifically in the context of FRTs, we nonetheless believe that such efforts are possible, important, and timely in today's environment.

APPENDIX A

Beyond Benchmarks and Datasets

As described in the introduction, lawmakers, technologists, and the general public are rapidly becoming aware of the risks and potential drawbacks of Facial Recognition Technologies (FRTs). Because modern FRTs are heavily influenced by the datasets used to develop and test them, many of the suggestions for improving these systems center around building new datasets. In this appendix, we discuss what datasets are, their role in the development and testing of FRTs, their shortcomings, how they could be improved, and importantly, some of the problems that datasets alone cannot solve.

To begin, we define two heavily used terms: *training set* and *benchmark*. For clarity, we will use definitions that are slightly narrower than those in technical articles, emphasizing how these definitions are used in FRTs.

Definition A.1

*A **training set** is a collection of images and associated labels that are used to help develop FRTs. For instance, a training set might consist of 100, 000 face images corresponding to 10, 000 different people. For a face recognition training set, each image would be labeled with a unique identifier of the person pictured, such as the person's full name. To use such a dataset in the development of an FRT, the parameters of the application are adjusted until it reports the correct name for as many of the images as possible. Intuitively, a training set can be thought of as "practice data" for an FRT system. At the end of this process, the application typically has the ability to correctly identify many of the people in the training set, even in pictures it has not been given before. However, for some pictures, it will still be unable to identify the person, or will give the wrong identity. An example of the types of images and labels (in this case, people's names) typically used in a training set are shown in Figure 7. Finally, it is important to understand that the training set is used when the application is being developed, not during a real deployment, which typically occurs at a later time. The set of people to be recognized at deployment time is called the **gallery**, and usually does not include any of the people from the training set.*

Because training sets are central to the development of many FRTs, a common solution proposed for various problems in FRT is to improve training sets. For example, if a training set contains few or no examples of faces of women, classifiers trained on this dataset will be optimized for faces of men. Such classifiers typically have lower accuracy identifying women. Building balanced training sets may reduce problems caused by these imbalances to some degree.

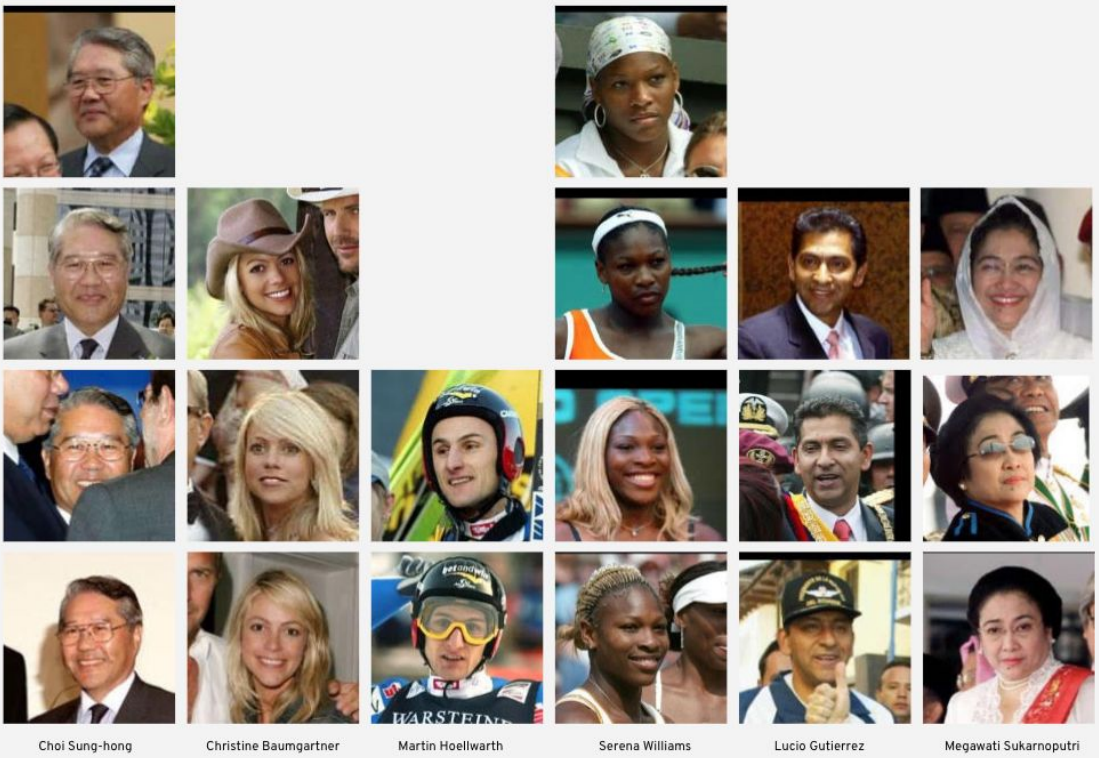

Figure 7: Part of a training set for face identification.

The figure shows what a portion of a training set for face identification might look like. The set of images shows images of six different people, and has between two and four images of each person.[4] Many modern training sets contain millions of face images and thousands of different identities.

Training sets are one important data source that support FRTs. The other major source is *benchmarks*.

[4] These images were taken from the Labeled Faces in the Wild [37] database. While such a database could theoretically be used as training data for an FRT system, it is not typically used for this purpose due to its relatively small size.

Definition A.2

*An FRT **benchmark** is a well-specified set of tests that are used to measure the performance of an FRT system for a particular problem. Benchmarks are used to test an FRT and give an estimate of its performance under certain conditions. It is important to understand that the performance of FRTs in one setting may or may not be related to their performance in another setting. For example, if a system scores well on a benchmark with high-resolution images (images with a great deal of detail), it may or may not perform well on low-resolution images.*

Benchmarks have played an important role in the research community, helping researchers understand which methods can improve accuracy in certain circumstances and which do not. However, as FRTs have moved from research labs into society at large, some benchmarks have been used to promote the efficacy of commercial facial recognition software in ways that were not intended. Benchmarks can help describe the performance of FRTs under certain conditions, but say little about how the system will perform when deployed under previously untested conditions.

There have been numerous **databases** of face images published throughout the years with increasing size and scope [51, 56, 66, 37, 43, 49]. Some have been used as a source of *training data* for developing FRT systems while others have been used mostly as *benchmarks* for evaluating such technologies. Still others are used in both regards. In some cases, authors do not explicitly state the purpose of a database, and practitioners use it as they see fit. In addition, the role of face databases might change over time as new and larger face databases are created. For example, the *Labeled Faces in the Wild* (LFW) database [37] was originally used as both a training set and as a benchmark. However, due to its relatively small size, it is now used almost exclusively as a benchmark, with developers using much larger datasets for training. Although the analysis in the rest of this appendix will focus on highlighting issues with benchmarks, some of these issues apply to any data collections including the training sets currently used for developing most FRTs.

The remainder of the appendix is structured as follows. In Section A.1, we explore the inability for benchmarks to capture deployment scenarios. As discussed in Section A.2, benchmarks can be overused, as when FRT systems are engineered to perform well on a specific benchmark. Interpreting metrics and how well they capture the performance of FRTs is discussed in Section A.3. Building benchmarks is itself a challenging task that requires careful consideration of ethical and accountability issues, as discussed in Section A.4. Finally, we discuss in Section A.5 how benchmark performance is not necessarily representative of FRTs that are currently used in deployment scenarios.

A.1 Issues with capturing deployment scenarios

Benchmarks for FRTs have had significant improvements in recent years in terms of both their size and diversity. However, they are still are limited, and are unable to capture all possible aspects of deployment. The FERET (Facial Recognition Technology) database, published in 1998, included 14,126 face images belonging to 1,199 individuals [56]. The benchmark's documentation however states: "For the evaluation procedure to produce meaningful results, the images in the developmental portion of the database must resemble those on which algorithms are to be tested. The development and testing datasets must be similar in both quality and quantity." Later databases for facial recognition often contain images compiled from images of faces on the internet. Newer benchmarks also contain disclaimers and warnings about the limitations of their benchmarks, including their limited representation of young people, babies, older adults, women, and many ethnicities. Documentation associated with benchmarks frequently emphasizes that a system's performance on the benchmark will say little about its performance in different conditions or on different populations.

Factors that affect the ability for a benchmark to cover every possible future deployment scenario include the resolution of images; the camera used to capture images; the lighting conditions when images were captured; whether faces in the images are forward-facing, in profile, or at some other angle; the size of the face in an image; the facial expression of faces in the images; the demographic distribution of people in the images including gender, race, and age; whether one or possibly multiple faces might be in any given image. This list is not exhaustive, and many other factors can affect the performance of FRTs. If a system has been tailored to produce good results on images from drivers' licenses or passports, then it will be unlikely to do well on images taken in other environments, such as outdoors, with more varied lighting conditions, poses, and facial expressions. Because a benchmark's usefulness is limited by its representation of different demographics, some benchmarks have been developed with specific target populations, such as the Japanese Female Facial Expression (JAFFE) database [46] or the Indian Face Database [39]. Generally, it is not possible to build a benchmark that covers and assesses the accuracy of a particular FRT for every possible population under arbitrary image conditions.

Challenge A.1

It is impractical to develop a comprehensive benchmark that covers arbitrary variations across many factors such as population demographics, image quality, pose, facial expressions, and camera viewpoint.

One ambitious goal would be to develop a more comprehensive database that exhibits diversity across a wide range of possible conditions. Unfortunately, building such a diverse benchmark is impractical for several reasons. First, it is very difficult to even enumerate all of the possible demographic groups a system might face in an arbitrary future scenario, let alone collect enough examples from each of those groups. Second, there are arbitrarily many ways in which the

images' conditions might change, from lighting conditions to image resolution, and many other factors. Even ensuring that a benchmark has diverse representation along only gender and race is challenging. Achieving such diversity requires obtaining sufficient samples from members of minority groups, obtaining permissions from large numbers of people, and addressing differing methods for defining racial diversity, including appearance, genetics, and self-identification. In order to test the performance of FRTs under more general conditions, attempts have been made to build more comprehensive face databases that exhibit variation along some of these axes. These benchmarks, while capturing a wider range of variations, may still have significant exclusions for a combination of testing conditions that could be seen during deployment.

A.2 Issues with benchmark overuse

Public benchmarks may cause manufacturers, either intentionally or not, to have algorithms that specialize on benchmark data at the expense of performing well on the 'real' data where the system is meant to be deployed. This is a phenomenon known as *overfitting*. If developers of FRTs have access to benchmark data, then a given system can be over-engineered by making repeated changes that directly improve its benchmark performance. This can be mitigated by designing benchmarks with *sequestered* data in which participating systems cannot access the test data directly. In such benchmarks, software must be submitted for evaluation to an independent third party in charge of the benchmark. In particular, the National Institute of Standards and Technology (NIST) has maintained a Face Recognition Vendor Test (FRVT) [14, 30] where participating teams must submit their FRT software to NIST for evaluation. In this case, the data is sequestered for multiple reasons. First, it contains images belonging to proprietary and governmental databases, including mugshots and visa photos. Second, NIST wants to ensure that FRT is not over-engineered to perform well on their benchmarks.

Academic benchmarks are also sometimes sequestered, though they generally are accompanied by training sets that were captured in similar conditions to the benchmark data. This allows designers to tailor their system's performance using detailed knowledge of the types of images that the sequestered benchmark is likely to contain. For instance, if a benchmark's corresponding training set contains only mugshot and visa-style photographs, a system designer can build a system to work well on these types of images. While such a system may perform well on the benchmark, there is no guarantee it would perform well in different conditions that are not represented in the benchmark, such as images of faces taken in uncontrolled environments.

Challenge A.2

Benchmarks become stale over repeated use. Over many interactions with a benchmark, developers may produce methods that do well on the benchmark but perform poorly in other scenarios.

Benchmarks become stale over repeated use. This is related to the overfitting issue, but may occur even when developers attempt to avoid overfitting. When participants are able to test their

systems on a benchmark, they can use the results to keep improving their models. After many repetitions, the system may over-specialize and produce accuracy numbers that are not representative of another scenario, even one that may be quite similar.

This issue is sometimes addressed by limiting the number of times each developer can use a benchmark. Even so, participants may learn from each other what type of components in their models lead to good numbers (through scientific publications, for example), thus partially bypassing the limitation on number of uses. Even with the best of intentions, a benchmark becomes less effective over time, due simply to the limited statistical power of a limited dataset. For this reason, benchmarks–even if intended for a deployment scenario that matches testing conditions–have a finite lifetime and should be updated over time. The process of updating or maintaining a benchmark may be as laborious and expensive as it was to produce the initial benchmark.

In summary, diverse benchmarks with strictly sequestered data may offer some improvements over prior benchmarks but should still not be taken as a general validation of the effectiveness of a method for arbitrary deployment conditions. Such benchmarks represent some advantages for applications that are intended for more than one deployment scenario in terms of measuring the expected accuracy of an FRT system. In the next section we discuss why reliance on accuracy metrics can also lead to different and problematic interpretations of benchmark reports.

A.3 Issues with benchmark metrics

Definition A.3

*A **metric** in the context of FRTs is a numerical measurement of how well or badly a system performs on a benchmark. Typical metrics include average error rates and average accuracy over the benchmark. Other common metrics include precision, recall, false positive rate, and false negative rate.*

Another issue with benchmarks is that a system's benchmark performance is usually quantified by some singular or small number of metrics and how these are interpreted. If a system is evaluated solely on its average accuracy on a benchmark, this will say little about the performance of the system in deployment, unless the deployment and benchmark have very similar capture conditions and demographics.

For example, consider the task of face verification. The most recent report by NIST [31] showed that some systems, when evaluated on two random pictures of different individuals, would identify the two different individuals as the same person (a false positive) on average less than one time per 10,000 trials. However, if the pairs of images were of people within the same country, sex, and age, this type of error happened more than three times in only 1,000 trials, more than thirty times as often as in the other experiment. If people are within the same country, sex, age, and regional location, false matches would likely be much more common than false matches for two arbitrary people across the globe. If face verification is used for individuals within the same demographic, then initial estimates of failure are likely to be severely

underestimated. This example demonstrates that the precise way systems measure their errors is critical. For this reason, studies often report more than one value of a single, aggregate metric, aiming to capture different factors that can affect the performance of a system.

Aggregated scores are especially problematic when sensitive or legally protected demographic variables such as gender, race, and age are considered. In particular, aggregated scores may conceal differences in performance within sub-groups. An FRT system that has substantially different performance on different groups can lead to significantly greater harms for one group than another.

There is ongoing research in how to build models that minimize differences in performance across sub-groups, but there has not yet been a completely satisfying solution to this problem. Moreover, disparate performance across sub-groups is often difficult to assess, as benchmarks often do not include demographic annotations. In this regard, the latest Face Recognition Vendor Test from NIST has analyzed performance across groups of people along gender and race. They found that many of the submitted algorithms underperformed for faces of African American, Asian and Native American individuals. However, each vendor's FRT system had their own weaknesses, and these differences depended on the particular facial recognition task. A recent NIST report [31] states "Since different algorithms perform better or worse in processing images of individuals in various demographics, policy makers, facial recognition system developers, and end users should be aware of these differences and use them to make decisions and to improve future performance." One example of such disparity provided by NIST was again for the face verification task where falsely matching two pictures as the same person occurred at a rate of 46 per million when both individuals were Polish, but at a rate of 2,400 per million when both individuals were Vietnamese.

Challenge A.3

There is a desire for a benchmark with a "universal checkbox" that, if checked, would declare software safe and effective. But such a benchmark must depend upon the specific details of every target deployment, and must incorporate this site-specific information. Our position is that there is no such universal checkbox.

A more general misuse of metrics can also happen when a restricted set of metrics become the objective of the underlying technology. Metrics, as emphasized earlier, summarize aspects of the systems they test, and as such; may fail to capture nuances of the underlying system that might affect its normal performance. An aggravating situation is when an FRT is tailored specifically to produce high numbers on these types of metrics, often at the expense of other considerations. Careful design and choice of metrics should ideally be dependent on the targeted deployment by taking into account differences among groups and other deployment specific variables of interest. In the field of economics this is know as Goodhart's law which is commonly stated as "When a measure becomes a target, it ceases to be a good measure." Ultimately a more responsible approach to measuring the quality of FRTs includes monitoring several metrics that measure diverse aspects of the underlying system and also continuous measurement and

tracking of metrics after deployment. In particular, it is important to avoid deliberate changes in the systems to artificially inflate the measurements obtained through established metrics.

A.4 Issues with accountability and consent

Since images of faces are closely related to the identity of people, there are several issues regarding accountability and consent that are difficult to overcome. We discuss here some of these aspects and the challenges in mitigating them.

Regarding accountability, benchmarks may be used for the development of a wide variety of products. Face matching and verification systems have possible uses from marketing and entertainment to applications in law enforcement. Each of these scenarios pose different requirements and risks, and no single benchmark is suited to consider all of these risks simultaneously. Furthermore, even if a benchmark considers the risks posed in such varied applications, they rarely prescribe how to interpret a system's benchmark performance, and what insights should be gained for different applications. A system's high performance on a benchmark may be used to justify using the system in some setting with very different deployment parameters than were used to design the benchmark, and the benchmark will generally say very little about how this system will perform under those different circumstances.

There are numerous real-world examples of this type of mismatch. In one case, Harvey and LaPlace [34] document a case of a video pedestrian detection benchmark that was later used to develop a facial recognition benchmark, even though the initial benchmark's intended purpose was entirely different. Documentation accompanying a benchmark is currently the best way to describe a benchmark's intended uses, but this does not restrict downstream users from applying a benchmark in some very different context. Benchmarks may then be used to make misleading claims about a system's performance; doing so makes it clear which parties are responsible when the system performs worse in other contexts. An additional concern stems from the fact that participants in a benchmark may provide consent for their likeness to be used in some ways and not others.

Challenge A.4

How can benchmarks be built that respect laws and privacy while being useful?

Issues of consent become more complicated when a benchmark might be extended for different uses. For instance, a benchmark for face verification could be augmented with user ratings of perceived physical attractiveness for the purpose of automatically finding "attractive" individuals. Individuals may have consented to using their likeness for the purpose of face matching, but they may not have consented to be used for determining a subjective measure of attractiveness that could affect them negatively or in unpredictable ways. Benchmarks and developers of benchmarks are often not able to control for these types of misuses by third

parties, nor is there an obvious mechanism for enforcing these third parties to adhere to the consent given by individuals in a benchmark. A recent proposal made by Gebru et al. [29] was to use datasheets that accompany datasets and include information about the data collection process, and its intended use. Perhaps a similar approach is needed for benchmarks where the additional consideration is on the way benchmark results can be used for commercial purposes.

Finally, building a benchmark is usually a challenging task which requires sourcing thousands or even millions of images. The larger the benchmark, the more reliable are the metrics and results obtained on the benchmark. An added challenge is building benchmarks large enough to be useful and getting appropriate consent from users. Current copyright practices often do not consider the use of images for developing FRTs. For example, Creative Commons licenses are mostly aimed at controlling re-use, authorship attribution, and re-distribution. It is unclear to what extent these licenses grant rights over their use to develop FRTs and what would be the right mechanisms to protect the rights of both the photographer and the subject depicted in the picture. Images of pictures in the public domain (e.g., images on the web) usually do not have a direct consent from the subjects depicted on the images. Mechanisms to build practical databases and better practices in benchmark creation are required.

There have been ideas in the community to perhaps attempt to use synthetic images of faces for building datasets and benchmarks that respect privacy and showcase diversity. While there has been significant progress in recent years in the domain of automatic synthesis of faces, there is still no comprehensive database or study that validates their effectiveness compared to the use of non-synthetic data. Moreover, successful methods for automatic face synthesis still need to be trained on non-synthetic images of faces. This leads to some of the same issues regarding privacy – as individual identities might leak in the synthesized faces – or diversity of the data – as the original diversity of the training data (or lack of it) might be replicated in the synthesized data.

A.5 Issues with the adoption and distribution of technology

Another issue in making decisions based on benchmark results is that they usually measure the latest research methodologies. These state-of-the-art methods may not be representative of what is available at the present time to end users. There is often a significant gap between the time a technology first becomes available until it reaches end users. This happens for various reasons: legal and technical constraints slow the distribution of updated software. Automatic updates of software mitigate these delays somewhat compared to historical updates, which took place via distribution of software on disks. However, there is still lag in upgrading systems, due to costs and the possibility that an upgrade might break a system which relies on the software.

Finally, once a technology has evidence of efficacy, many challenges remain in bringing the technology to common users. For instance, the technology might rely on computational resources that are not available to end users, and more work may be needed to adapt the technology to consumer devices. If the technology will be deployed as a resource “in the cloud”, then there will be significant time to scale such solutions to thousands of simultaneous users. Therefore, recent results on a benchmark do not necessarily reflect the performance of FRTs currently in use. Moreover, developers might choose to deploy a version of their system that is

not the most accurate but instead the one that has the best compromise between accuracy and speed.

Finally, we remark that the performance of the best FRTs on a benchmark do not represent the performance of all FRTs. Not all companies rely on the same type of systems, nor do all companies have the same resources to develop their technical solutions, or build training datasets that are diverse and maximally useful. This difference in resource availability can mean that the top performers do not represent the accuracy of most commercially available software. Moreover, according to the most recent facial recognition vendor tests by NIST, no single company seems to have the monopoly on the best system across all tests and benchmarks. Therefore, consideration of the appropriateness of FRTs for a given situation is not only dependent on the task and application but also on the specific technical solution being evaluated as opposed to the best possible technical solution possible for a given problem.

For instance, in the task of facial identification, the latest vendor test by NIST reports that with good quality portrait-style photos, matching the picture of an individual against a database containing pictures for 12 million individuals is highly accurate. The closest matching image will be a false positive match only one in a thousand times for the most accurate algorithms in existence today. However for other algorithms submitted to the benchmark the top matching image will be a false positive almost half of the time. The FRVT report remarks that "This large accuracy range is consistent with the buyer-beware maxim, and indicates that face recognition software is far from being commoditized."

A.6 Summary of benchmark discussion

Benchmarks are a powerful tool for developing and improving current and existing FRTs given the success of data-driven machine learning methods¸ However their use is problematic for assessing FRTs for deployment in user facing applications. The issues discussed in this appendix include the relationship between training and test distributions, the lack of diversity in current benchmarks and the difficulty in aiming to capture a diverse range of scenarios, the problems with relying on metrics to determine accuracy and the issues arising from consent considerations and accountability. In addition, it is worth noting that while many advances have been made in FRTs, there is a wide range of software that is currently available in this area and the methods that perform best are not representative of all commercially available FRTs. This raises questions such as what are the required levels of accuracy and error that are tolerable for a given application.

References

[1] Best practices for common uses of facial recognition technologies. *Federal Trade Commission*, October, 2012.

[2] Row over AI that 'identifies gay faces'. *BBC*, September 11, 2017. https://bbc.com/news/technology-41188560.

[3] Facial recognition technology: Fundamental rights considerations in the context of law enforcement. Technical report, European Union Agency for Fundamental Rights, November 21, 2019.

[4] Recommendation of the council on artificial intelligence. Technical Report OECD Legal 0449, Organization for Economic Co-operation and Development, May 21, 2019.

[5] The history of FDA's fight for consumer protection and public health. https://fda.gov/about-fda/history-fdas-fight-consumer-protection-and-public-health, 2020.

[6] Researchers said their 'unbiased' facial recognition could identify potential future criminals — then deleted the announcement after backlash. *Business Insider*, May 7, 2020.

[7] An Act Concerning the use of Facial Recognition Services. *2020 Regular Session of The Washington State 66th Legislature*, Senate Bill 6280, March, 2020.

[8] Administrative Code - Acquisition of Surveillance Technology. *City of San Francisco, CA*, May, 2019.

[9] Ban on Town Use of Face Surveillance Technology. *Town of Brookline, MA. Town Hall Meeting*, Article 25, November, 2019.

[10] Law enforcement: Facial Recognition and other Biometric Surveillance. *California Penal Code 832.19*, Assembly Bill No. 1215. Chapter 579, October, 2019.

[11] Mark Andrejevic and Neil Selwyn. Facial recognition technology in schools: Critical questions and concerns. *Learning, Media, and Technology*, November 5, 2019.

[12] J. Michael Bailey, Paul L. Vasey, Lisa M. Diamond, S. Marc Breedlove, Eric Vilain, and Marc Epprecht. Sexual orientation, controversy, and science. *Psychological Science in the Public Interest*, 17(2):45–101, 2016.

[13] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pol lak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[14] Duane M Blackburn, Mike Bone, and P. Jonathon Phillips. Face recognition vendor test 2000: Evaluation report. Technical report, Defense Advanced Research Projects Agency. Arlington, VA, 2001.

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc., 2016.

[16] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[17] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[18] The Electronic Privacy Information Center. In the Matter of HireVue, Inc.: Complaint and Request for Investigation, Injunction, and Other Relief. November 6, 2019. `https:// epic .org/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf.`

[19] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jeremy L. Tipton, and Arun R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2018.

[20] Kate Crawford. Regulate facial-recognition technology. *Nature*, 572(1):32–41, August 29, 2019.

[21] Sen. Christopher A. Coons [D-DE] and Sen. Mike Lee [R-UT]. Facial Recognition Technology Warrant Act of 2019. *In the Senate of the United States*, 116th Congress, 1st Session. S.2878, November 14, 2019.

[22] Sen. Cory A. Booker [D-NJ]. No Biometric Barriers to Housing Act of 2019. *In the Senate of the United States*, 116th Congress, 1st Session. S.2689, October 23, 2019.

[23] Sen. Jeff Merkley [D-OR] and Sen. Cory A. Booker [D-NJ]. Ethical Use of Facial Recognition Act. *In the Senate of the United States*, 116th Congress, 2d Session. S.3284, February 12, 2020.

[24] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Technical Report 2020-1, Berkman Klein Center, 2020.

[25] Yasmin Gagne. How we fought our landlord's secretive plan for facial recognition—and won. *Fast Company*, November 22, 2019. https://fastcompany.com/90431686/our-landlord-wants-to-install-facial-recognition-in-our-homes-but-were-fighting-back.

[26] Clare Garvie. America under watch: Face surveillance in the United States. *Georgetown Law, Center on Privacy & Technology*, 2019.

[27] Clare Garvie. Garbage in, garbage out: Face recognition on flawed data. *Georgetown Law, Center on Privacy & Technology*, 2019.

[28] Clare Garvie, Alvaro M. Bedoya, and Jonathan Frankle. The perpetual line-up: Unregulated police face recognition in America. *Georgetown Law, Center on Privacy & Technology*, 2016.

[29] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[30] Patrick J. Grother, Mei L. Ngan, and Kayee K. Hanaoka. Face recognition vendor test (FRVT) Part 2: Identification. *National Institute of Standards and Technology*, 2019.

[31] Patrick J. Grother, Mei L. Ngan, and Kayee K. Hanaoka. Face recognition vendor test (FRVT) Part 3: Demographic effects. *National Institute of Standards and Technology*, 2019.

[32] Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B. Kamphausen, Martin Zenker, Lynne M. Bird, and Karen W. Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25:60–64, 2019.

[33] Isobel A. Hamilton. Google suspended facial recognition research for the Pixel 4 smartphone after reportedly targeting homeless black people. *Business Insider*, October 17, 2019. https://businessinsider.com/google-suspends-facial-recognition-research-after-daily-news-report-2019-10.

[34] Adam Harvey and Jules LaPlace. MegaPixels: Origins, ethics, and privacy implications of publicly available face recognition image datasets. https://megapixels.cc, 2019.

[35] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018.

[36] Paddy Hillyard and Steve Tombs. Beyond criminology?. *Beyond criminology: Taking harm seriously*, 13, 2004.

[37] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[38] Natalie Jacewicz. Why are health studies so white? *The Atlantic*, June 16, 2016. https://theatlantic.com/health/archive/2016/06/why-are-health-studies-so-white/487046/.

[39] Vidit Jain and Amitabha Mukherjee. The Indian Face Database. http://vis-cs.umass.edu/vidit/IndianFaceDatabase/index.html, 2002.

[40] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1:389–399, September 2, 2019.

[41] George Joseph and Kenneth Lippe. IBM used NYPD surveillance footage to develop technology that lets police search by skin color. *The Intercept*, September 6, 2018. https://theintercept.com/2018/09/06/nypd-surveillance-camera-skin-tone-search/.

[42] Zolan Kanno-Youngs and David E. Sanger. Border agency's images of travelers stolen in hack. *The New York Times*, June 10, 2019. https://nytimes.com/2019/06/10/us/politics/ customs-data-breach.html.

[43] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The Megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[44] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[45] Erik Learned-Miller, Qifeng Lu, Angela Paisley, Peter Trainer, Volker Blanz, Katrin Dedden, and Ralph E. Miller. Detecting acromegaly: Screening for disease with a morphable model. In *Medical Image Computing and Computer-Assisted Intervention*, volume 2, pages 495–503, 2006.

[46] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. The Japanese female facial expression (JAFFE) database. In *Proceedings of Third International Conference on Automatic Face and Gesture Recognition*, pages 14–16, 1998.

[47] Nicole Martinez-Martin. What are important ethical implications of using facial recognition technology in health care? *AMA Journal of Ethics*, 21(2):E180–E187, February 1, 2019.

[48] Lucas Ramon Mendos. State-sponsored homophobia 2019. *International Lesbian, Gay, Bisexual, Trans and Intersex Association*, 2019.

[49] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[50] Ralph E. Miller, Erik Learned-Miller, Peter Trainer, Angela Paisley, and Volker Blanz. Early diagnosis of acromegaly: Computers vs clinicians. *Clinical Endocrinology*, 75:226–231, 2011.

[51] Baback Moghaddam and Alexander P. Pentland. Face recognition using view-based and modu- lar eigenspaces. In *Automatic Systems for the Identification and Inspection of Humans*, volume 2277, pages 12–21. International Society for Optics and Photonics, 1994.

[52] Mei L. Ngan and Patrick J. Grother. *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. National Institute of Standards and Technology, 2015.

[53] Sam S. Oh, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E. Bercelo, Marquitta J. White, Danielle M. de Bruin, Ruth M. Greenblatt, Kirsten Bibbins-Domingo, Alan H. B. Wu, Luisa N. Borrell, Chris Gunter, Neil R. Powe, and Esteban G. Burchard. Diversity in clinical and biomedical research: A promise yet to be fulfilled. *PLOS Medicine*, 12(12), 2015.

[54] Kari Paul. 'Ban this technology': Students protest US universities' use of facial recognition. *The Guardian*, March 2, 2020. `https://theguardian.com/us-news/2020/mar/02/facial-recognition-us-colleges-ucla-ban.`

[55] Gisela Perez and Hilary Cook. Google, YouTube, Venmo and LinkedIn send cease-and-desist letters to facial recognition app that helps law enforcement. *CBS New*, February 5, 2020. `https://cbsnews.com/news/clearview-ai-google-youtube-send-cease-and-desist-letter-to-facial-recognition-app/.`

[56] P. Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[57] Jon Porter. Facebook and LinkedIn are latest to demand Clearview stop scraping images for facial recognition tech. *The Verge*, February 6, 2020. https://theverge.com/2020/2/6/21126063/facebook-clearview-ai-image-scraping-facial-recognition-database-terms-of-service-twitter-youtube

[58] Sen. Roy Blunt [R-MO] and Sen. Brian Schatz [D-HI]. Commercial Facial Recognition Privacy Act of 2019. *In the Senate of the United States*, 116th Congress, 1st Session. S.847, March 14, 2019.

[59] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, page 246–257, 2019.

[60] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, page 145–151, 2020.

[61] Dan Reilly. Musicians and fans unite to keep facial recognition tech out of concerts. *Fortune*, September 30, 2019. https://fortune.com/2019/09/30/ban-facial-recognition-live-events-music-festivals-concerts/.

[62] Joshua R.Scannell. This is not minority report. In Ruha Benjamin, editor, *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life.* Duke University Press.

[63] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.

[64] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 59–68. ACM, 2019.

[65] Evan Selinger and Woodrow Hartzog. The inconsentability of facial surveillance. *66 Loyola Law Review 101*, pages 1–22, March 19, 2020.

[66] Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression (PIE) database. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 53–58. IEEE, 2002.

[67] Brad Smith. Facial recognition: It's time for action. https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/, last accessed on May 1, 2020.

[68] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. *Association for Computational Linguistics (ACL)*, 2019.

[69] Luke Stark. Facial recognition is the plutonium of AI. *Crossroads: THe ACM Magazine for Students*, 25(3):50–55, April, 2019.

[70] Josh Taylor. Major breach found in biometrics system used by bank, UK police and defence firms . *The Guardian*, August 14, 2019. https://theguardian.com/technology/2019/aug/14/police major-breach-found-in-biometrics-system-used-by-banks-uk-police-and-defence-firms.

[71] Janet M. Torpy, Cassio Lynm, and Robert M. Golub. Hyperthyroidism. *JAMA*, 306(3):330–330, 07 2011.

[72] Alina Tugend. Fervor grows for regulating big tech. *The New York Times*, November 11, 2019. https://nytimes.com/2019/11/11/business/dealbook/regulating-big-tech-companies.html.

[73] Harrisburg University. Facial recognition software predicts criminality, researchers say. *Communications of the ACM*, May 6, 2020. https://cacm.acm.org/careers/244713-facial-recognition-software-predicts-criminality-researchers-say/fulltext.

[74] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordóñez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *International Conference on Computer Vision (ICCV)*, October 2019.

[75] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2):246–257, 2019.

[76] Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *CoRR*, abs/1611.04135, 2016.

[77] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.

[78] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.