JULY 2021

Facing the Risk

Part 1: Mapping the Human Rights Risks in the Development of Facial Recognition Technology

AUTHORS
Amy K. Lehr
William Crumpler

A Report of the CSIS Strategic Technologies Program and the Human Rights Initiative

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

JULY 2021

Facing the Risk

Part 1: Mapping the Human Rights Risks in the Development of Facial Recognition Technology

AUTHORS
Amy K. Lehr
William Crumpler

A Report of the CSIS Strategic Technologies Program and the Human Rights Initiative

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

About CSIS

The Center for Strategic and International Studies (CSIS) is a bipartisan, nonprofit policy research organization dedicated to advancing practical ideas to address the world's greatest challenges.

Thomas J. Pritzker was named chairman of the CSIS Board of Trustees in 2015, succeeding former U.S. senator Sam Nunn (D-GA). Founded in 1962, CSIS is led by John J. Hamre, who has served as president and chief executive officer since 2000.

CSIS's purpose is to define the future of national security. We are guided by a distinct set of values—nonpartisanship, independent thought, innovative thinking, cross-disciplinary scholarship, integrity and professionalism, and talent development. CSIS's values work in concert toward the goal of making real-world impact.

CSIS scholars bring their policy expertise, judgment, and robust networks to their research, analysis, and recommendations. We organize conferences, publish, lecture, and make media appearances that aim to increase the knowledge, awareness, and salience of policy issues with relevant stakeholders and the interested public.

CSIS has impact when our research helps to inform the decisionmaking of key policymakers and the thinking of key influencers. We work toward a vision of a safer and more prosperous world.

CSIS does not take specific policy positions; accordingly, all views expressed herein should be understood to be solely those of the author(s).

Acknowledgments

CSIS thanks William Carter, Mariefaye Bechrakis, Anna Lehman-Ludwig, and Luiza Parolin for their research and administrative support during the drafting of this report. CSIS also thanks the nearly 100 individuals who participated in interviews and workshops over the course of this project.

This report was made possible through a grant from the U.S. State Department Bureau of Democracy, Human Rights, and Labor.

Center for Strategic & International Studies
1616 Rhode Island Avenue, NW
Washington, D.C. 20036
202-887-0200 | www.csis.org

Contents

Executive Summary

The development process for facial recognition technologies (FRTs) has three primary stages: training data collection, algorithm development, and software integration. At each stage, there is an opportunity for actors to either contribute to, or mitigate, the human rights risks associated with the system's eventual operation.

- **Training data collectors** are responsible for collecting, labeling, and curating the databases of face images that developers use to train facial recognition systems. Training data collectors can impact individuals' privacy rights depending on where and under what conditions they source their images. Collectors can also violate subjects' right to remedy if they do not provide the means to discover whether a person is contained within a data set and request removal. Finally, collectors can contribute to discrimination risks if they are not deliberate and transparent about the demographic composition of the images they are compiling.

- **Algorithm developers** are responsible for building the facial recognition algorithm itself by selecting an appropriate neural network architecture and training it on large quantities of face images. This work may lead to discriminatory effects when the system is deployed if insufficient attention is given to the training data being used. Inadequate testing of demographic effects could also lead to facial recognition systems that have higher error rates for certain groups of people. Additionally, if an algorithm consistently demonstrates high error rates in certain situations, and algorithm developers are either unaware or not transparent about these limitations, it could lead to risks of misidentification once the system has been deployed.

- **Software integrators** incorporate trained facial recognition models into applications and platforms and sell the finished systems to operators. Integrators can affect human rights primarily through

their decisions about which operators to sell to and how to monitor the way their systems are used. Integrators can also reduce potential rights impacts by providing training to help ensure that operators use the system in a responsible way and by designing their products to automatically minimize data collection and protect against misuse and abuse by operators.

In some cases, a single developer is responsible for all three stages in the development process, while in others the work is spread out among separate firms that contract with one another. The structure of the FRT supply chain can take a variety of different forms, with three of the most common being illustrated in this report through fictionalized case studies:

- **End-to-End Provider:** A single company has responsibility for the full development chain of an FRT product, which is provided directly to operators to install and run locally.
- **Facial Recognition as a Service (FRaaS):** A single company develops an FRT product that it then makes available to others as a service through the cloud.
- **OEM-ing Facial Recognition:** Rather than sell directly to operators, algorithm developers sell to intermediary hardware and software providers who integrate facial recognition capabilities into their own offerings.

A variety of fundamental rights and freedoms are affected by the decisions made by the actors involved in FRT development.

- **Right to Privacy:** Depending on how training data is collected, the privacy rights of the individuals included in training data sets may be impacted, especially if images are used without the subjects' awareness or consent. While it is unlikely that the FRT industry will pivot to consent-based collection practices in the near future, there are some sources of training data that pose greater risks than others, and collectors should strive to minimize privacy risks to the extent possible. Privacy risks can also emerge once the system is deployed if software integrators do not practice privacy by design during the development process. Relevant practices for integrators may include ensuring that FRT systems encrypt and segregate sensitive data, minimize data collection and storage, enforce retention limits, and have high default thresholds for similarity scoring.
- **Right to Non-Discrimination:** A majority of facial recognition systems have been shown to have different accuracy rates for different demographic groups, raising the risk that some groups may be disproportionately impacted by errors. Training data collectors can contribute to this risk if they do not take steps to ensure that the data they collect is broadly representative of the population that the system is used on. Algorithm developers can contribute to this risk if they do not test their algorithms to determine whether bias is present. And software integrators can contribute to non-discrimination if they sell FRT systems to operators who use the technology in harmful and discriminatory ways.
- **Right to Effective Remedy:** Training data collectors do not always provide individuals the means to determine whether their information is being used to train FRT algorithms or to request removal from training data sets. This restricts an individual's right to challenge practices that impact their privacy rights. Algorithm developers and software integrators can also infringe on this right to the extent that they fail to provide grievance mechanisms for individuals affected by their systems once they have been sold to operators.

- **Other Fundamental Rights and Freedoms:** Through their decisions of who they will and will not sell their products to, software integrators can have significant impacts on a large number of rights, including freedom of expression, movement, assembly, and expression; freedom from arbitrary arrest and detention; and the rights to privacy, non-discrimination, and life, liberty, and security. Integrators can mitigate these risks by instituting screening mechanisms to determine whether potential customers are at high risk of contributing to human rights abuses and by providing operators with comprehensive training to reduce the chances of harm caused by operator error.

This report concludes by considering the steps that different actors in the FRT value chain can put in place to address their potential human rights impacts, drawing on the UN Guiding Principles on Business and Human Rights as a framework. The report first identifies efforts that extend across all firms involved in the development process and then provides more specific recommendations for particular actors according to the role they play in the supply chain.

General Recommendations for Firms

1. Provide a policy statement outlining human rights commitments.
2. Assess the human rights impacts of products and services.
3. Integrate and act on impact findings.
4. Track the effectiveness of mitigation efforts.
5. Communicate how impacts are being addressed.
6. Provide opportunity for remedy.

Recommendations for Training Data Collectors

1. Proactively share information about the source, demographic composition, and other details of training data sets with algorithm developers.
2. Assess privacy implications prior to the assembly of a training data product.
3. Establish policies of data security and collection minimization to reduce the risks of unauthorized access.
4. Establish systems to allow individuals to determine whether they are included as part of a training data set and to request removal when applicable.
5. Conduct due diligence on potential buyers of training data to assess the risk that the models being developed could lead to human rights abuses.
6. Develop contractual clauses limiting the use of the training data.
7. Practice transparency about company policies and practices relating to data collection and undergo regular auditing to provide independent verification of compliance with these policies.

Recommendations for Algorithm Developers

1. Evaluate any training data sourced from outside providers to ensure it is demographically representative of the population likely to be affected by the algorithm in development.
2. Rigorously test the performance of facial recognition models to determine their accuracy and identify whether any demographic biases are present.
3. Conduct due diligence on any software integrators that models are licensed to and include language in contracts and licensing agreements that would allow the developer to sever ties if evidence of abuse emerges.
4. Develop contractual clauses limiting the use of the algorithm.
5. Communicate human rights policies and practices.

Recommendations for Software Integrators

1. Perform rigorous testing on models purchased from outside developers to determine their accuracy and ensure they are free from demographic biases.
2. Institute internal structures and processes during the development process for identifying and escalating the potential human rights concerns of staff relating to new products and services.
3. Establish external advisory bodies with representatives from a wide range of disciplines to provide outside assessment of the propriety of potentially risky new products or services.
4. Practice principles of privacy and data protection by design and default.
5. Conduct due diligence of potential buyers to assess the risk of deployments leading to human rights abuses.
6. Leverage contractual or other mechanisms to establish processes for controlling or regularly reviewing how customers are using the tools provided to them.
7. Provide rigorous and accessible training for customers to help operators understand how to use the technology in ways that respect human rights.
8. Practice transparency about company policies and practices relating to product development and sale, and undergo regular auditing to provide independent verification of compliance with these policies.

In addition to improving the policies and practices of individual actors in the facial recognition supply chain, it will also be necessary for government and industry consortia to push suppliers to adopt stronger human rights safeguards. This work should focus on the following themes:

- **Improving Accuracy and Eliminating Bias:** Transparency requirements, certification bodies, and the adoption of common methods of testing and reporting for facial recognition performance can help level the playing field and reward developers who build accurate and unbiased systems.
- **Privacy, Notice, and Consent:** Nations must have strong domestic legal frameworks clarifying standards for notice and consent when collecting biometric information for facial recognition

training. They must also establish rights of access, correction, and erasure for individuals included in training databases.

- **Human Rights Due Diligence:** Policymakers can grant clarity to integrators about how to conduct due diligence assessments of potential customers and what uses may be impermissible.
- **Remedy:** Governments should ensure there is sufficient transparency to enable individuals impacted by FRT systems to seek remedy for any harms they experience and adequate enforcement mechanisms to ensure that firms are adhering to expectations regarding accuracy, bias, and data protection.

1

Introduction

Thanks to a decade of rapid progress in the field of computer vision, facial recognition technology (FRT) has become a commercial product available to almost any government or business in the world. Organizations ranging from law enforcement agencies to independent retail outlets are beginning to explore ways of integrating FRT into their operations. Proponents hope that facial recognition may support public safety initiatives and improve access to services, but the risk of errors and abuse means that FRT deployments carry substantial risks to a variety of fundamental rights and freedoms. As facial recognition systems continue to grow cheaper, more powerful, and easier to use, these risks will only grow. All of this is only possible thanks to the ecosystem of developers whose work has brought these FRT tools onto the market.

Though developers are not the ones determining how and when FRT is used, they nonetheless bear responsibility for ensuring that their products and services do not end up contributing to violations of fundamental rights and freedoms through failure or abuse. Because this technology has only emerged recently, its ecosystem of developers and suppliers is still poorly understood by those working on the risks of surveillance technologies. This report seeks to help illuminate the structure of this industry and examine the ways that FRT developers can contribute to or mitigate human rights impacts through their decisions. Using the UN Guiding Principles on Business and Human Rights (UNGPs) as a guide, the report presents a set of recommendations for each group of actors in the FRT supply chain to demonstrate how firms can change their business practices to ensure they respect internationally recognized human rights.

A companion report, *Facing the Risk Part 2: Mapping the Human Rights Risks in the Deployment of Facial Recognition Technology*, provides a complementary analysis of the risks involved in the deployment and operation of facial recognition systems.

2

Stages of FRT Development

The development process for facial recognition begins with the collection of face images to use as training data. In this stage, data scientists and artificial intelligence (AI) researchers collect, curate, and label thousands or even millions of face images. The next stage is algorithm development, which involves selecting an appropriate neural network architecture and training it using the collected face data to generate a facial recognition model. Once the facial recognition model has been trained, software integrators incorporate these functions into apps or platforms. After this development process is finished, facial recognition systems are deployed by operators who access the software either by installing it locally on their own systems or by signing up for a cloud platform that provides facial recognition capabilities as a service.

A large variety of actors and sectors are involved in the FRT development process, often playing more than one role. Some companies have the capability to accomplish all of these steps in-house, whereas others may specialize in one particular area and then sell their products or services to other actors further down the supply chain. This means that firms may have different opportunities and obligations to consider potential rights impacts depending on where they sit within this chain and which other actors they work with. The following sections outline the three primary stages of facial recognition development in greater detail.

Stage 1: Training Data Collection

Facial recognition developers depend on access to large amounts of high-quality training data to build their models.[1] Many major tech companies compile their own proprietary training data sets using images they collect as data controllers. One example of this is Facebook taking advantage of its millions of user-uploaded photos to train its FRT systems for photo tagging. Another example is Paravision, which began as the photo sorting app Ever before pivoting to use the photos they had collected to build facial recognition tools.[2]

FRT developers without their own source of user data may instead decide to enter into contracts or form partnerships with other firms. The AI firm Clarifai, for example, trained its facial recognition system using photos from the dating site OkCupid and an undisclosed social media site.[3] This kind of sale and reuse of face data is usually allowed under the terms of service agreed to by users of these sites. The study team also heard through interviews that some developers have made use of government passport and visa photo databases to refine their algorithms, though so far there has been little public reporting on how frequently this occurs.

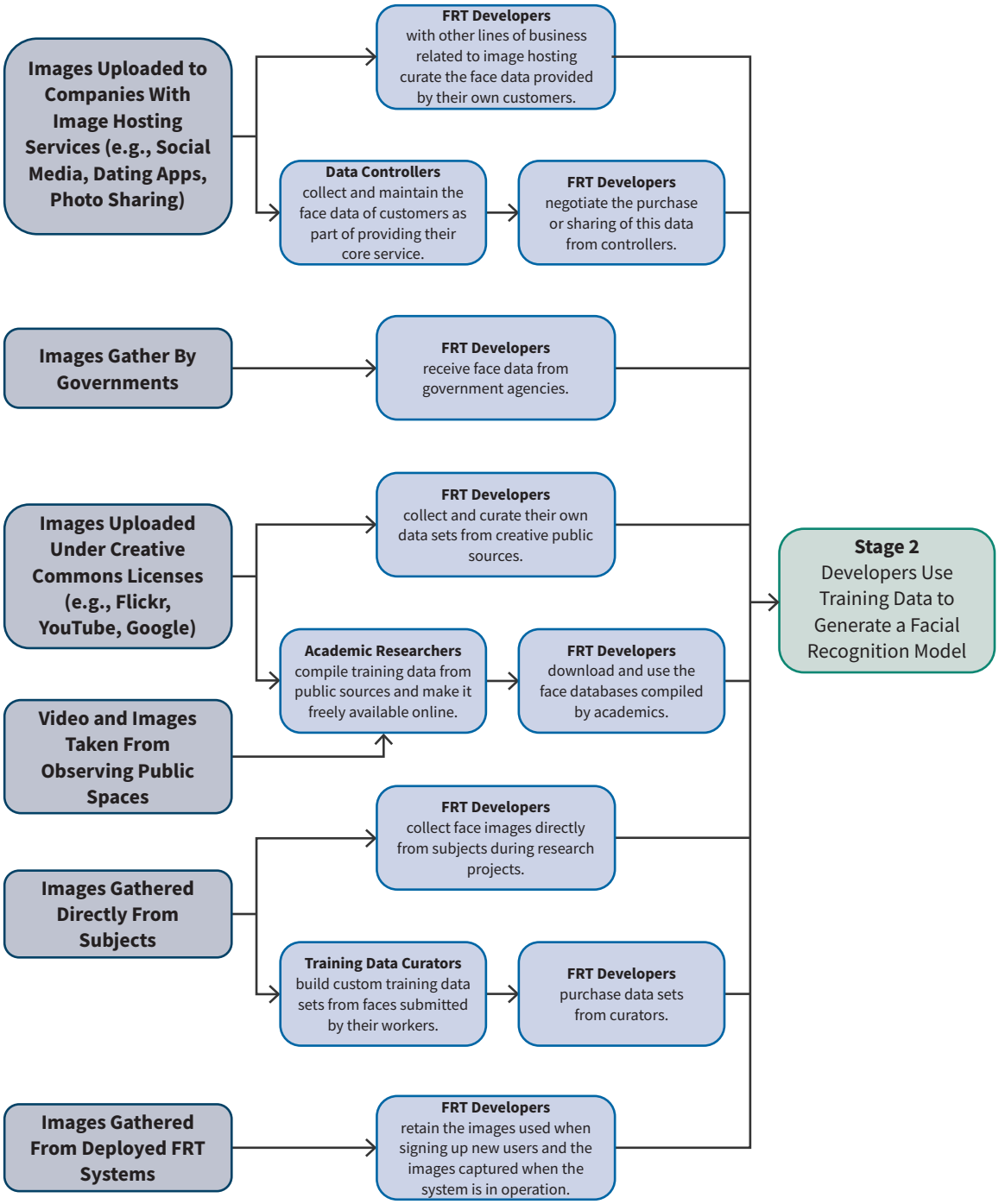
Academic researchers and smaller FRT developers without partnership opportunities or access to the kind of data held by giants such as Facebook may instead source images from websites such as Flickr, YouTube, or Google, where certain photos or videos may be publicly accessible under creative commons licenses.[4] Some university researchers have also collected face samples from footage of public places, such as college campuses or coffeeshops.[5] Almost every data set created by academic researchers, and even some created by corporate researchers, is published openly to encourage other researchers to benchmark their own algorithms against them. Most of these data sets are intended only for academic research, where they serve the purpose of encouraging common standards for facial recognition accuracy and can improve researchers' access to more diverse data sets to help reduce algorithmic bias.[6] However, a number of such data sets also allow the data to be used for commercial purposes, and even those that do not have few safeguards to prevent exploitation.[7]

It is also possible to build an original training database by going directly to individuals and getting their consent to take their picture for the specific purpose of training a facial recognition system. It is generally not feasible for this to be the only source of face data due to the time and cost of gathering the large quantities required, but developers sometimes use this strategy to gather smaller, high-quality data sets to support specific research projects to improve their products.[8] This kind of data collection can occur independently by companies or in partnership with university researchers.

There is also an emerging trend of private companies such as Lionbridge and ClickWorker that specialize in the custom sourcing and labeling of training data for facial recognition developers.[9] These firms often gather their data by having their own workers submit face images of themselves according to the exact specifications of the customer. The emergence of professionalized firms for gathering training data may help to bring rigor to the process—aiding both technical outcomes (e.g., accuracy and bias) as well as rights outcomes (e.g., respect for privacy)—but it could also result in training data sourcing becoming even more opaque as developers lose the ability to directly oversee collection procedures.

Finally, some firms will use the faces captured after the system is deployed to retrain their software. The Chinese firm CloudWalk, for example, signed a deal with the government of Zimbabwe in 2018 that included provisions allowing CloudWalk to use the data it gathered on African faces to improve

its algorithm's performance on subjects with dark skin tones.[10] While interviews for this report suggest that the Zimbabwe government has yet to actually deploy these systems in the country, these types of arrangements are likely to become more common in the future as the technology becomes more widespread.



Stage 2: Algorithm Development

For an in-depth discussion of how facial recognition algorithms function, readers are encouraged to reference CSIS's recent report on the subject, *How Does Facial Recognition Work?*[11] In brief, the vast majority of modern facial recognition algorithms use computer-generated filters to transform face images into numerical expressions (usually referred to as "references" or "templates") that can be compared to determine their similarity. Developers use AI to automate a process of trial and error that helps identify the best filters for generating these references.

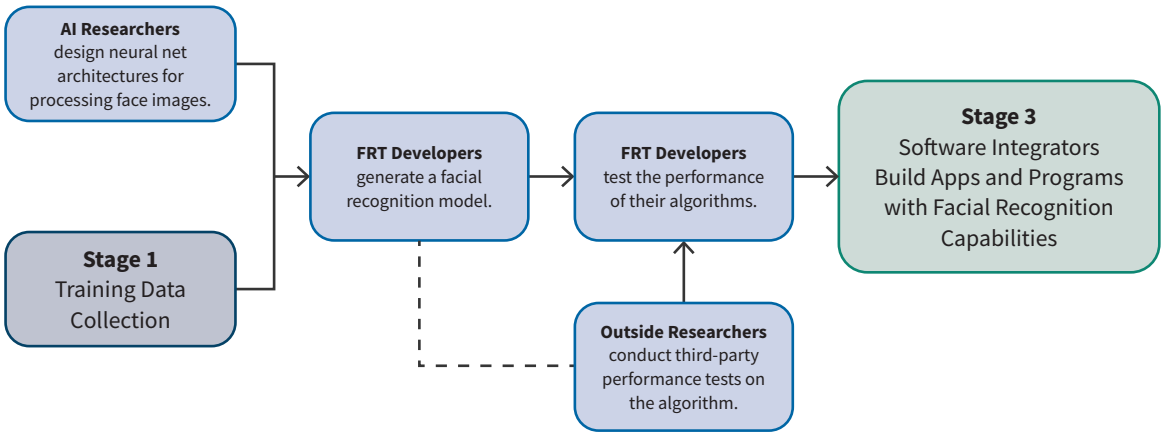
The neural net architectures used to structure this process are developed by AI researchers in university or corporate labs. Neural network design is highly specialized, and not every facial recognition developer has a team devoted to it. Many improvements in the field are published in scholarly journals, allowing others to take advantage of their discoveries. Enough information is freely available to enable tech firms without in-house AI researchers to still develop their own models, though major companies such as Google or Facebook with their own AI research divisions have an advantage in creating the most accurate systems.[12]

Once a neural network architecture has been selected, the system is trained by providing it with a series of "triplets"—collections of three face images where two of the faces belong to one person and the third belongs to someone else. The system turns each of the three images into a reference and then compares their similarity. The system is given the goal of achieving the maximum similarity for the references coming from the same subject and the minimum similarity for the references coming from different subjects.

As the system churns through tens or hundreds of thousands of these triplets, the algorithm continuously tweaks the operations its filters perform and then measures whether the changes it made result in better or worse accuracy in correctly determining which of the three images are from the same person. If a change leads to an improvement, the system keeps it, and if performance gets worse, the system will revert to its previous state and then try something else. In this way, the system slowly learns which filters are the best at creating distinctive face references. At the end of this process, the system arrives at a set of filters that have repeatedly proven their efficacy.

This helps explain why many facial recognition systems have been found to have different accuracy rates for different demographic groups.[13] If an algorithm is trained on a data set of mostly white males, the program will spend more time optimizing to recognize their faces and less time optimizing for other demographics. This can be avoided by selecting a representative training data set to begin with or by choosing data sampling techniques and algorithm architectures that are able to somewhat mitigate the negative effects of biased training data.[14]

After training, models are packaged together with other, similar systems that perform complementary tasks such as detecting when an image contains a live face and processing that face into a standardized format. The facial recognition systems are then tested to determine their performance. These tests can range from very basic technical evaluations to see how well the system performs on curated images to more detailed testing regimes that attempt to measure performance across different operational scenarios and different demographic groups. Some teams may also make their models available to outside researchers to conduct their own assessments.



Stage 3: Software Integration

Once the facial recognition models have been trained, they must be integrated into applications and platforms that allow end-users to deploy the systems. Software integrators are the ones who decide the kinds of tools and applications facial recognition models become a part of. This can include everything from the face verification programs used to unlock electronic devices to large-scale surveillance platforms. Integrators are responsible for packaging the facial recognition model together with other software and hardware, designing the user interface, configuring the system according to the intended use, and making it available to the end user.

In some cases, the software integrator is also the same company that developed the algorithm. Many major companies, such as Amazon in the United States, NEC in Japan, IDEMIA in France, and Hikvision in China, fit this profile. These companies both design the algorithms and build the software and hardware platforms and services that are sold to customers.

In other cases, integrators acquire facial recognition models from external sources, either by taking advantage of the fact that some developers have released pretrained facial recognition models online as part of open-source machine learning software libraries or by purchasing access to models developed by dedicated facial recognition developers.[15] One example of the latter case is DataWorks, which provides access to facial recognition tools through the mugshot management software it sells to law enforcement agencies.[16] DataWorks does not actually develop facial recognition in-house but rather subcontracts that out to firms such as NEC and Rank One, which develop their own facial recognition models and provide them to DataWorks.

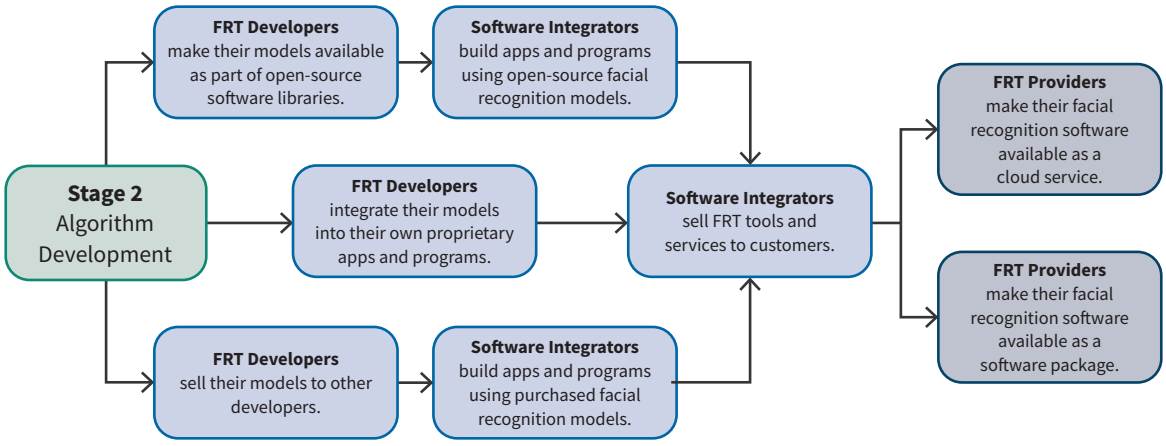
Other examples of this subcontracting can be found in the access control industry, where facial recognition aims to replace badging as a way to manage employee and visitor access to buildings.[17] Additionally, some security and surveillance camera manufacturers—especially at the low end—prefer to contract with facial recognition developers rather than develop algorithms themselves. This is particularly prominent in China, where the trend of "OEM-ing facial recognition"—selling software that customers can use and brand as their own—has created a proliferation of low-cost integrators reselling technology from developers such as Hikvision and Dahua.[18]

Regardless of how the responsibilities for software integration and algorithm development are distributed, at the end of the process integrators have two options for making their facial recognition systems available to customers. The first is to sell a software package that operators can install and run locally on their own systems. The second is to offer facial recognition as a service through a cloud platform.

The first option tends to be favored by most large enterprise and government users. It involves the software integrator working with the operator to install and configure the system so that the customer can use it independently. This typifies many types of law enforcement deployments, where the software is deployed through a direct partnership between the customer and developer (NEC is one prominent example of a developer with many such contracts).[19] Through interviews, the authors also found evidence of these types of deployments in the case of live facial recognition used by Indonesia's national police. These types of deployments ensure that data is only ever processed on the customer's network, as opposed to being uploaded to the developer for remote processing. This is usually the preference for most major deployments by enterprises and governments due to privacy and data protection concerns.

The second option is for operators to contract with providers offering facial recognition as a service (FRaaS). FRaaS providers do not sell their software to others. Instead, operators set up a system that sends images to the provider for analysis. The provider analyzes the faces on their end and then sends the result—either "match" or "no match"—back to the operator. This setup may be preferable to customers without the technical infrastructure for local solutions, as the highest-quality facial recognition matching often requires significant processing power in the form of GPUs, which may be prohibitively expensive to maintain for many small companies. A prominent example of this kind of hosted setup is Amazon's Rekognition, where customers can access Amazon's cloud-based facial recognition system through an online application program interface (API) to create custom watchlists and review matches from on-premises video streams.

Depending on the size of the match list and the complexity of the deployment, some FRaaS providers may have a more direct hand in assisting with the system setup than others. In general, however, companies offering FRaaS are more removed from the details of the deployments than companies that provide software directly to the operator.



3

Common Models of FRT Development

The following three fictionalized case studies are designed to help clarify the most common models of FRT development and how the decisions made at each juncture can change the overall impact of the system.

Case Study 1: IdentiT (End-to-End Provider)

IdentiT is a medium-sized tech firm that specializes in providing a variety of products and services focused on biometric identity verification and video surveillance. In the past, the company has had success providing its customers with fingerprint readers, iris scanners, security cameras, and associated software platforms. Its customers have included government border control agencies as well as major corporations looking for access management solutions and security systems. Responding to growing interest by its customers in adding facial recognition as an alternate biometric, IdentiT decides to develop a new facial recognition product line that will allow it to take advantage of its in-house expertise with biometric verification, software development, and hardware design.

IdentiT collects no images through its other lines of business, so its development team is forced to look outside the company for data to train their algorithm. The company strikes a deal with two local app developers who collect photos through their consumer apps. These developers compile curated lists of their users' photos, strip them of identifying data, and then make them available for IdentiT to use when training their algorithm. These apps' terms of service allow user data to be sold and shared in this way, though the large majority of users are likely unaware of this. These databases give IdentiT access to sufficiently large quantities of faces to build a powerful facial recognition algorithm, though because the companies are primarily active in just one region, the final training data set is highly skewed toward the demographic profile of IdentiT's home country.

IdentiT's data scientists proceed to generate a facial recognition model by applying this training data to a neural network architecture adapted from recent academic literature. IdentiT conducts a basic suite of internal technical evaluations to determine the algorithm's performance. However, these are limited to tests involving the comparison of still images purchased from the two local app developers. IdentiT does not evaluate the system's performance in real-world conditions. However, the company does submit its algorithm for testing by the U.S. National Institute of Standards and Technology (NIST) and is ranked in the top 25 percent of algorithms across a variety of tests.

IdentiT's first potential customers are three border control agencies who are interested in deploying the system at airports and border crossings to improve the convenience and security of their traveler checks. Early in the negotiation process, IdentiT decides to look into these organizations to determine the risks that their systems would be used or repurposed in ways that could pose risks to the travelers processed by the system. Their analysis includes evaluating the stated purpose and scope of the agencies' planned deployments, the legal and human rights context of the countries they are located in, and any previous instances where the agencies had been linked to abuses. Through this evaluation, the company decides to exclude one of the prospective agencies, whose government has been complicit in human rights abuses in the past and has a legal system that offers few guarantees that the information being collected will not contribute to a mass surveillance regime.

For the two other agencies, IdentiT proceeds to provide them with the camera hardware necessary to run the facial recognition system, as well as a software package that will allow them to independently enroll individuals into the program and verify travelers at border checkpoints. IdentiT sends an employee to each location where the system is being installed to conduct in-person trainings over several days to help the staff understand not only know how to set up and operate the system, but also how to properly interpret match results, identify the potential sources and risks of errors during operation, and understand how to properly handle and secure the data that is collected.

The first to deploy the system is the border control agency of the country where IdentiT is headquartered. This deployment proceeds smoothly. Because the border checks take place in a relatively controlled setting, the accuracy of the system is reasonably high, and the agency's decision to have travelers simultaneously verify themselves via their fingerprints helps provide a safeguard against the risk of accidentally letting the wrong person through. Because the workers manning the stations have been trained on the system, they are able to capably adjudicate any errors quickly with minimal disruption. Inspired by this upgrade, an airline decides to trial the technology as a way of passively checking in their passengers as they enter through the boarding gate. However, the airline finds that this unconstrained setting results in too many false positives and pauses use of the technology once they realize it would be more costly to deal with the bookkeeping mistakes of these errors than it would be to check passengers in the traditional way.

The second border control agency to deploy the system does so in a different region of the world than the one where IdentiT has gathered its training data. Because of this, they experience a greater number of errors compared to the first agency, which is exacerbated by the fact that they do not use an additional biometric modality such as fingerprints to perform a second check of travelers. The operators are reasonably happy with the technology, however, despite two incidents in the first several months of operation when a traveler is falsely identified as high risk and detained for several hours while the authorities verify their identities.

Two years after the initial deployment, an IdentiT support team that is visiting to help provide training and support for a new update finds that the second border agency has configured the system so that traveler logs are being compiled and sent to the nation's law enforcement and intelligence services. This is taking place without public knowledge or any obvious set of guardrails to prevent abuse. Because the software package has already been delivered, IdentiT does not have the ability to take away the agency's access to the technology but does decide to cease providing updates or support for the system.

Case Study 2: Pulsar (Facial Recognition-as-a-Service)

Pulsar is a major multinational digital giant that started as a social media company but has expanded into various other lines of business ranging from e-commerce to cloud and enterprise software. Pulsar began developing image analysis and face recognition software to improve the tagging of faces and images in its social media and e-commerce services but recently decided to take advantage of its advanced software by offering cloud-based facial recognition-as-a-service (FRaaS) to its enterprise and government customers.

Thanks to its position as an established social media platform, Pulsar sources training data from its own users. The company does not gain explicit prior consent from its users, as the reuse of user-provided data to improve the platform is already covered under the terms of service. Because of Pulsar's global reach, its data scientists can collect and label millions of user photos from a variety of demographic groups around the world.

The algorithm generated from this data benefits from the work of Pulsar's in-house AI researchers and undergoes extensive testing for accuracy and bias on both still images and simulated operational environments. Pulsar also creates an application program interface (API) that allows third parties to apply to conduct their own assessments of the algorithm. By working with academic researchers, the company is able to identify potential areas of concern they had not picked up on before and improve the system's performance. The company also submits its algorithm for testing by NIST and is ranked in the top 10 percent of algorithms worldwide across a variety of tests.

Pulsar does not sell access to the software itself but instead creates a web app that allows others to submit images or footage to be remotely processed. Organizations with systems already in use for security monitoring, access management, and a wide variety of other purposes are able to integrate Pulsar's cloud service to roll out facial recognition without any of the overhead required to independently deploy and maintain a new software program. This offering is particularly attractive to small and medium customers without the expertise or technical infrastructure to deploy a facial recognition program on-site.

One of Pulsar's first clients is a factory which institutes facial recognition as a way to clock its workers in and out. They view facial recognition as a way to address issues they have been experiencing with workers sharing the cards formerly used for attendance. The factory uploads the faces of their workers to the cloud and sets up a camera so that workers can automatically be verified at a station as they enter and exit work. Retail stores also sign on and use the system to upload pictures of individuals who have been caught shoplifting and to run their security camera streams through Pulsar's service to send an alert to a manager whenever one of these individuals enters a store.

When signing up for the service, Pulsar requires its clients to provide basic identifying information, but for most customers, it does not follow up to verify information or investigate their intention for the software before granting them access. This is primarily due to the large expense of conducting assessments for each one of the many customers Pulsar supports. As a result, in a majority of instances the company has little insight into the types of projects and organizations it is serving. The primary exception to this is the case of large customers, who often require more extensive support from Pulsar when setting up their infrastructure. These customers have more regular contact with Pulsar technical support and sales teams in ways that grant Pulsar some degree of insight into the scope and purpose of deployment.

Pulsar works with the government of the country where they are headquartered to receive information about agencies and companies around the world that are under investigation or are suspected of being involved in human rights abuses. This has allowed Pulsar to implement a high-level screen to identify whenever one of these high-risk organizations attempts to sign up and deny them access to the platform.

For large customers, Pulsar often has support teams that assist the customer in learning how to operate the system. For its smaller customers, the company develops documentation and an online training platform that it encourages operators to go through. The training materials primarily focus on how to operate the system, and while it includes some information about how to interpret results and configure the settings to avoid errors, the completion of these sections is the responsibility of the operator.

Pulsar does not monitor what its customers upload. This is both due to a concern that the company could invite heavier regulatory scrutiny if it began taking a more active role in how its customers are processing personal data, and because the resulting fears over security, privacy, and business confidentiality could lead many of Pulsar's clients to take their business elsewhere.

As a result, Pulsar is not aware when several of the stores it works with begin to expand their watchlists to include the photos of suspects provided by the local police force. The stores are being asked by police to alert them when any of these suspects are detected, as the police are not authorized to run real-time facial recognition themselves using government cameras. It is only after local reporting breaks the story and an NGO raises the issue with Pulsar that they investigate the claims and decide to cut off access to the stores that have been implicated.

Case Study 3: QSP (OEM-ing Facial Recognition)

QSP is a small security integrator that specializes in selling video surveillance systems. QSP provides low- to mid-range equipment to customers which lack the funds for state-of-the-art devices and software. Historically, QSP had focused on traditional security cameras, but growing interest from their customers in facial recognition capabilities makes the company feel that they need to add this capability to their suite of products or lose ground to competitors.

QSP is not responsible for developing the algorithm used in their products. Instead, they purchase specialized chipsets and software from larger firms that have built their own facial recognition capabilities. The supplier providing the software is referred to as the original equipment manufacturer (OEM), a common term for instances where a company sells their products to another firm who rebrands and resells them as their own.

As the company is not involved in the software's development, QSP has little information about the quality of the facial recognition algorithm. When searching for companies to source from, few are transparent about the details of their model or the data used to train it. All offer similar sales pitches claiming to have over 99 percent accuracy, but none point to any trusted, standardized assessments that prove their claims. QSP lacks the technical capacity to conduct assessments themselves, so they make their final decision primarily as a result of price considerations.

As a low-cost provider with thin margins, QSP has never implemented a process for vetting its customers to determine whether they may use the technology in ways that could be harmful. QSP sells its technology to anyone that is willing to buy, unless the company is aware of the potential for legal repercussions in a particular sale. This is rare, as the country where QSP is headquartered does not actively work with technology companies to help them understand the current risks and targets of legal measures.

QSP's clients include a wide variety of organizations, from small retail outlets to law enforcement agencies. QSP's technology is provided directly to its customers to run and operate independently, leaving them little insight into how the systems are used once they have been sold. This lack of visibility is compounded by the fact that QSP provides only nominal follow-up support to their clients—a result of them only having a superficial understanding of the software they have purchased from an outside firm. This also means that the training QSP provides to their clients is limited and mostly just repackaged from the documentation provided to them from the firm from which they sourced their facial recognition model.

One of the first buyers of QSP's new facial recognition-enabled product line is a law enforcement agency that is eager to install cameras at a local market to monitor for wanted criminals. QSP completes the sale of a dozen facial recognition-enabled cameras along with the associated software. Due to a lack of support, training, and familiarity with the technology, 10 of the cameras are installed at heights and angles that lead to consistently poor-quality footage. Paired with the relatively low-quality algorithm being used, these cameras give matches that are so frequently and obviously incorrect that the police department effectively abandons their use.

The two remaining cameras, however, do manage to capture some usable matches, and the police use this system to arrest a number of individuals by sending alerts to officers on the ground with a description of the person who has been detected. However, the overall accuracy rate is still quite low, and in a large number of cases, the police falsely detain individuals due to mistaken alerts by the facial recognition system or due to officers with incomplete descriptions approaching the wrong individual. The majority of these cases are resolved during the stop when the officer checks the suspect's ID, but in several instances the police do not discover they have arrested the wrong person until several days after taking the suspect into custody. In a handful of instances, the system allows the police to identify and approach suspects wanted for serious crimes, but it also leads to the arrest of several people for minor offenses and, in one case, the arrest of a protester wanted by the national government for their role in a recent demonstration against the ruling party.

4

The Human Rights Impacts of Facial Recognition Development

This section identifies some of the potential impacts of the various actors in the FRT value chain. It is not exhaustive. Indeed, an exhaustive list is not feasible because FRT leads to risks that can affect a broad range of rights, depending on the purpose for which FRT is deployed.

The Right to Privacy

TRAINING DATA COLLECTION

Depending on how training data is collected, the privacy rights of the individuals included in training data sets could be impacted. The training of facial recognition systems requires data scientists to assemble huge databases of thousands or millions of face images. In the majority of these cases, images are taken and used for training without the subjects' awareness or consent. This is both for the sake of more easily gathering the required quantities of data and for improving the accuracy of the resulting model by ensuring that the facial recognition systems are trained on the types of images they are likely to encounter when in use—those featuring unaware subjects in public settings.

While explicit consent would be the ideal standard for justifying the collection and use of this data, this is unlikely to ever become common practice in the industry due to the sheer quantity of data required for these systems. Of the alternatives to databases created with subject consent, there are some options which pose greater risks than others. In particular, the practice of government agencies sharing publicly managed photo databases with developers poses significant privacy risks due to the involuntary nature of government databases and the violation of citizens' expectations that the data they provide to their government will not be shared with private enterprises. The collection of data from operational systems also creates significant risks if users are unaware when they enroll in the system that their data will be used for more than just the operation of the platform.

The practice of sourcing data from outside firms also raises privacy concerns. Though this sharing may be allowed under the terms of service that users agree to, the nature of this sharing is often opaque to the user. Even after signing terms of use agreements, most remain uncertain of when and for what purposes their information can be transferred to other companies.

While it does not eliminate privacy concerns, the practice of sourcing training data from a developer's own users or from images posted publicly online under creative commons licenses creates comparatively fewer privacy risks. In the case of companies which train on their own users' images, use is more likely (though by no means guaranteed) to be aligned with the subjects' privacy expectations and to minimize the number of organizations that have access to individuals' personal data. In the case of images scraped from the web, the risks are reduced because the people who uploaded the photos made the choice to do so under open licenses, even though they had the option to place stricter limits on how their images could be used.

While explicit consent would be the ideal standard for justifying the collection and use of this data, this is unlikely to ever become common practice in the industry due to the sheer quantity of data required for these systems.

It is also important to note where training data collection does not impact privacy rights. There is a common misconception that individuals included in training databases can later be recognized by that system once it has been deployed in the field. This is incorrect. The only individuals a system is capable of identifying are those in the system's match list—a database of faces compiled by the end user (rather than the developer) after the model has already been trained and deployed. In the example of an access control system used by a business, the match list would be the set of employees who had signed up to allow the building's facial recognition system to automatically grant them entry after a face scan. Inclusion in a training data set is not the same as inclusion in a match list. However, if a person included in the system's training data is also included in the match list, there is some evidence to suggest that the system will be slightly more reliable at identifying them correctly compared to someone that the system was never trained on.[20] Depending on the context, this could have either positive or negative implications for the individual.

SOFTWARE INTEGRATION

Privacy by design is an approach to software development that emphasizes the importance of building technological systems in ways that automatically protect data from being accessed or used in unauthorized ways. Software integrators have an opportunity to reduce the privacy risks of their systems by incorporating this principle into their development process. Practicing privacy by design does not guarantee that a system will not violate subjects' privacy once deployed, but the following examples show how design choices can influence the risk profile of systems once they are put into use.

The most basic privacy protection a facial recognition system can employ is encryption. Unencrypted data is an easy target for hackers. Once unencrypted data is stolen, there is nothing stopping others

from taking advantage of it. By making sure that all data collected by the system is stored in an encrypted form, the integrator can ensure that even if there were a data breach, it would be impossible to actually access and read any of the information that was taken. Encrypting all personal data and biometric templates generated by a system is a simple way to significantly reduce privacy risks.

Another way to reduce the risk of data breaches is to segregate personal and biometric information. If biometric information, such as an image or facial recognition template, is stored in the same location as personal information, such as a person's name, address, financial history, or medical information, it is more likely that a single data breach would give a malicious actor enough information to take advantage of the victims. However, if different types of sensitive data were stored separately, a single breach would be less likely to result in enough data being lost that a malicious actor could do a significant amount of harm.

Practicing privacy by design does not guarantee that a system will not violate subjects' privacy once deployed, but design choices can influence the risk profile of systems once they are put into use.

Privacy by design also includes minimizing data collection and storage. If a piece of data is not necessary for the system to perform its tasks, it should not be kept by the operator. For example, integrators can design their systems to automatically delete any images or templates collected from people not already enrolled in the system. For example, a casino may wish to install a system that scans each person as they enter to determine whether they are on a list of known gambling addicts and thus should be removed from the premises. This will involve capturing the image of a large number of visitors who are not on the watchlist and who the casino has no cause to retain any information about. By having the system set to automatically delete the images and templates of those people once they have been confirmed to not be on any watchlist, the facial recognition system can reduce the risk of privacy violations caused by overbroad data collection.

Minimization should also include deleting the raw images used to enroll users. When a person signs up for a facial recognition system, their photo is taken and used to generate a template. Whenever the system wants to do a comparison between that subject and another face, it compares the templates it generates, not the images themselves. This means that the photos taken of people's faces are not necessary once a template has been generated. Storing these images even after a template has been generated creates unnecessary privacy risks, which can be addressed by setting the system to automatically erase images once they are no longer needed.

Systems can also be set to automatically delete or prompt operators to review facial recognition templates and other data files after a certain amount of time. If a person signs up for a particular facial recognition service, such as a store's customer loyalty program, but then moves away or stops using that service, there is nothing that keeps the operator from simply holding onto their data indefinitely. Integrators can build their systems to adhere to a certain time-limited retention policy automatically

or prompt operators to review data that may no longer be needed and delete those profiles manually. This reduces the risks created from data breaches, by minimizing the amount of information an operator would have at any given time and helps ensure that operators do not have the opportunity to continue processing data after the purpose of its collection has been fulfilled.

Finally, integrators can set the default similarity threshold of the system in a way that minimizes the risks stemming from misidentification. Similarity thresholds are the standard facial recognition systems used to determine whether a given pair of faces should be returned as a match. The system estimates a particular similarity score between the images and then compares that to the threshold score set by the operator (or, if the operator has not set a threshold, the default put in place by the integrator). If the similarity score calculated for a given pair of faces is above this threshold, it returns a match.

Where these thresholds are set has important implications for the impact of the system once it is deployed. Systems using a low similarity threshold are more likely to return false positives—instances where the system mistakenly believes two people are the same. This can lead to negative consequences when facial recognition is used in contexts such as law enforcement. There is no universal answer to the question of what similarity threshold facial recognition systems should use. It is highly context and vendor dependent. In theory, each operator would receive training about how to set a threshold appropriate to their use case. However, in practice, operators tend to simply use the system in its default setting without modifying its parameters. This means that the defaults set by integrators can have a large impact on the real-world consequences of deployments.

The Right to Non-Discrimination

TRAINING DATA COLLECTION

The Universal Declaration on Human Rights (UDHR), International Covenant on Civil and Political Rights (ICCPR), and International Covenant on Economic, Social, and Cultural Rights (ICESCR) all guarantee individuals the right to non-discrimination, and this right cannot be restricted. Facial recognition presents risks to the principle of non-discrimination if systems have different accuracy for members of different demographics. These disparities can have significant consequences for a broad array of rights further along the value chain by causing more errors for certain groups.

The presence of demographic bias in facial recognition performance is well documented. The most thorough investigation of demographic effects in facial recognition accuracy was completed by the U.S. National Institute of Standards and Technology (NIST) in 2019. Through their testing, NIST confirmed that a majority of algorithms exhibit demographic differences in both false negative rates and false positive rates.[21]

Using curated images, NIST found that, in general, Asians, African Americans, and American Indians had higher false positive error rates than white individuals, women had higher false positive rates than men, and children and the elderly had higher false positive rates than middle-aged adults. Differences in false positive rates are generally of greater concern, as there is usually greater risk in misidentifying someone than in having someone be incorrectly rejected by a facial recognition system. NIST found that demographic factors had a much larger effect on false positive rates—where differences in the error rate between demographic groups could vary by a factor of 10 or even 100—than false negative rates—where differences were generally within a factor of 3.

The most important factor in determining whether a facial recognition system exhibits demographic bias is the training data used to generate the model. Because algorithms learn to identify people based on the examples they are fed, training data sets that provide few examples of particular groups often lead to algorithms that are less accurate at recognizing people from those groups once they are deployed in the real world. Recent research has shown that many of the most prominent training data sets currently in use are heavily skewed toward Caucasian males, which helps explain facial recognition algorithms tend to perform better on this group.[22]

Demographic bias can be reduced if training data collectors are more rigorous in collecting broadly representative data sets. For example, NIST found that algorithms developed in China performed better on Asian faces than algorithms developed in the United States—a difference likely attributable to the fact that the training data sets used by Chinese developers contained a greater representation of Asian faces. It is particularly important for algorithm developers to ensure that the training data is representative of the population that will be using the completed system. One Asian facial recognition developer the authors interviewed noted that they had observed a 10 to 20 percent differential in error rates between models that were trained on local faces before a deployment and those that were not. Encouragingly, NIST's analysis found that some of the algorithms they tested demonstrated no detectable demographic biases, indicating that the issue is addressable given better training data selection and ongoing improvements to the underlying technology.

It is worth noting the inherent tension between the right to non-discrimination and the right to privacy in the context of training data collection. Rectifying the underrepresentation of certain demographic groups in training data sets requires collecting additional image sets featuring those groups. Prioritizing privacy interests during training data collection would make it more difficult and more expensive for this information to be collected.

Because algorithms learn to identify people based on the examples they are fed, training data sets that provide few examples of particular groups often lead to algorithms that are less accurate at recognizing people from those groups once they are deployed in the real world.

ALGORITHM DEVELOPMENT

Choices made during algorithm development can also impact the ultimate risks of the system. Developers that take advantage of the latest findings and methods of AI researchers are more likely to build accurate systems than those relying on older techniques. Testing has demonstrated that the most accurate algorithms overall also tend to have the smallest demographic biases.[23] Developers also have the opportunity to choose data sampling techniques and algorithm architectures that are able to somewhat mitigate the negative effects of biased training data.[24] Developers who do not take advantage of these tools and techniques are more likely to produce algorithms that have discriminatory effects.

Developers can also reduce the risk of discriminatory effects by implementing rigorous testing regimes to assess the performance of their algorithms. All developers perform some kind of testing on their algorithms, but some spend more time and resources on the process than others. At the most basic level, testing involves separating out some of the training data before generating the model, and then testing the completed algorithm on that data to measure its performance. This provides some information about the error rate of the algorithm but does not necessarily tell the developer anything about potential bias and cannot reveal if the nature of the training data source itself may have introduced blind spots.

Better testing regimes ensure that the images used for testing are similar to what the algorithm would likely see in an operational environment and include dedicated analysis to compare the error rates for different demographic groups. The most comprehensive testing regimes involve simulations where volunteers allow their photos to be taken in a setting that mimics how the system will eventually be used. These are expensive and difficult to scale, however, and therefore rarely take place. Developers also have the opportunity to open their algorithms for testing by third-party groups, which can help identify issues that were not surfaced during internal tests. Independent audits have already been shown to be an effective mechanism for catalyzing improvements in algorithmic bias.[25]

Given that the right to non-discrimination is absolute, there are questions as to whether FRT should be used at all in contexts where it can significantly affect the enjoyment of rights if this potential bias has not been addressed at this stage in the value chain.

Better testing regimes ensure that the images used for testing are similar to what the algorithm would likely see in an operational environment and include dedicated analysis to compare the error rates for different demographic groups.

SOFTWARE INTEGRATION

Depending on the quality of the training data and algorithms that a software integrator uses, the integrator may create a product that could have discriminatory effects and impact the enjoyment of other rights. Integrators who conduct due diligence on the origin and performance of the software and data that their system is based on can mitigate these risks.

Integrators also have a responsibility to ensure their products are not being sold to operators who will use them in discriminatory ways. Multiple Chinese facial recognition providers, for example, have offered products to Chinese authorities that would allow them to single out Uyghurs.[26] While this kind of "Uyghur detection" is actually a form of face characterization (given a single face, classify it according to certain variables) rather than face recognition (given two faces, determine the similarity between them), many operators will likely try to pair the two kinds of tools together, leading to broader technological systems that could both classify and identify an individual for discriminatory purposes.

The Right to Effective Remedy

TRAINING DATA COLLECTION

The ability of individuals to opt out of training databases is currently limited, and in most cases, subjects are not aware if they have been included in the first place. Some database managers allow individuals to request removal from training data sets, but the process is often cumbersome, and it is not always easy to determine whether a given person is present. Others do not provide the option to opt-out at all. This is a difficult issue to solve, as making it easier to determine whether an individual is included in a database would require linking more identifying data to the photos. Face images collected for training and testing are not usually linked to identifying information such as names or ID numbers. This helps to minimize the potential for abuse, as it is extremely difficult for someone who gains unauthorized access to the data set to identify the people included without already knowing their faces. However, it also means it is very challenging to inform individuals that they were included in a training database.

ALGORITHM DEVELOPMENT AND SOFTWARE INTEGRATION

If developers and integrators fail to take steps to identify, mitigate, and communicate the risks of misidentification or discriminatory effects of their products and services, these actors may end up either contributing to or being directly linked to adverse human rights impacts. These actors may also contribute to impacts if they do not institute adequate safeguards regarding the customers they sell to. In these instances, business enterprises should establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely affected. If firms fail to provide these mechanisms or are not transparent about their products and the identities of their customers, impacted groups and individuals will have a much more challenging time seeking remedy for the harms they have suffered.

Other Fundamental Rights and Freedoms

SOFTWARE INTEGRATION

Through their decisions of who they will and will not sell their products to, software integrators can have significant impacts on a large number of rights, including but not limited to freedom of expression, movement, assembly, and expression; freedom from arbitrary arrest and detention; and the rights to privacy, non-discrimination, and life, liberty, and security. Developers who are indiscriminate in the customers they sell to can contribute to human rights abuses, such as the oppression of Uyghurs in Xinjiang, China; the arrest and detention of protesters by governments; and the spread of mass surveillance in authoritarian countries. In some cases, there are opportunities for facial recognition providers to engage in some form of ongoing review of how their technology is used, allowing developers to identify and address abuses early. However, when software is simply being provided to an operator for them to set up and use themselves, the opportunities for oversight are limited, and the initial, pre-sale due diligence is the only opportunity providers have to identify potential risks.

Some of the risks of facial recognition emerge from users who are unfamiliar with how to use the systems properly. In particular, many operators overestimate the accuracy of the models, leading to the risk that false matches may result in action being taken against individuals without the operator taking the time to validate the result of a facial recognition match. This risk is highest in the context of

law enforcement, where some departments may attempt to arrest and detain individuals based solely on matches from facial recognition, despite the fact that these systems remain too inaccurate to serve as a reliable standard of justification for arrest. However, similar risks could apply to misidentification of shoplifters or other private sector contexts.

Developers who are indiscriminate in the customers they sell to can contribute to human rights abuses, such as the oppression of Uyghurs in Xinjiang, China; the arrest and detention of protesters by governments; and the spread of mass surveillance in authoritarian countries.

Providers of facial recognition systems have an important impact on these risks through the actions they take to educate the users of their technology. In some cases, facial recognition providers will work closely with the buyers of their technology to ensure they are fully trained to operate the systems. One Asian developer the authors interviewed had adopted the practice of sending the product manager out to the site of installation to conduct a week-long training with the operators prior to deployment. In other cases, however, facial recognition operators receive little more than a link to a few pages online as their training. These organizations often have no other exposure to individuals or organizations that could help them work through non-technical questions of how they should deploy their systems and are at much higher risk of misusing the technology.

5

Company Policies and Procedures to Address the Human Rights Impacts of FRT

This section considers the steps that different actors in the FRT value chain can put in place to address their potential human rights impacts, drawing on the UN Guiding Principles on Business and Human Rights (UNGPs) as a framework.[27] It first identifies efforts that extend across all firms involved in the development process and then provides more specific recommendations for particular actors according to the role they play in the supply chain.[28]

According to the UNGPs, companies have a responsibility to respect internationally recognized human rights. They do so by exercising human rights due diligence—having in place effective policies and procedures to identify and address potential and actual human rights impacts throughout their value chain. Due diligence steps include assessing actual and potential human rights impacts, integrating and acting upon the findings, tracking responses, and communicating how impacts are addressed.

How companies are expected to respond to impacts will vary depending on their relationship to the impact. If they cause an impact, they are expected to cease or prevent it. If they contribute to an impact, they should cease or prevent their contribution and use their leverage to mitigate its effects to the greatest extent possible. Businesses should also seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products, or services by their business relationships, even if they have not contributed to those impacts. They should use their leverage with their business partners to accomplish this. If they lack the leverage to prevent or mitigate adverse impacts and cannot increase their leverage, they should consider ending the relationship. Businesses also have a responsibility to provide effective remedies for human rights harms associated with their products and services.

This paper does not provide guidance regarding when a company involved in facial recognition development would be considered to cause or contribute to an impact versus when it would be

considered directly linked. However, it is noteworthy that some facial recognition issues such as bias could have significant ramifications several stages away in the supply chain, and multiple actors could bear some form of responsibility for any human rights impacts that arise.

The corporate responsibility to respect human rights is independent of whether governments are enforcing human rights-compliant laws and may in some cases require companies to adhere to higher standards than those set by national law. Given that national laws governing facial recognition directly or indirectly are often non-existent or nascent, this responsibility is particularly important. Both the assessment of human rights risks and the measures taken in response must account for this local context.

The relationship between the corporate responsibility to respect human rights and the existing human rights obligations of governments is closely interwoven. The UNGPs reiterate long-standing international law, noting that governments have a duty to protect human rights from adverse impacts by third parties, such as companies. This means they should have in place laws, regulations, enforcement, and remedy mechanisms dealing with those private sector actors involved in FRT development.

The UNGPs are a useful tool for considering the human rights responsibilities of companies that develop FRT and the policies and procedures they should have in place. They help companies establish appropriate internal systems and external engagement that consider the risks to human rights that products can cause when deployed by other actors or in particular circumstances. Many sectors risk being associated with human rights abuses through their suppliers. The technology sector faces not only this issue but also challenges related to how end-users deploy their products. This challenge is why both human rights by design and evaluation of customers and context are vital tools for the sector, as discussed below.

The UNGPs are a useful tool for considering the human rights responsibilities of companies that develop FRT and the policies and procedures they should have in place.

General Recommendations

Policy Statement: Every stage of the FRT value chain can be linked to potential adverse human rights impacts. Therefore, actors at each stage should have in place a policy statement outlining their human rights commitments. The statement of commitment should clearly set out the business's expectations for its personnel, business partners, suppliers, customers, and other linked parties. These expectations and commitments should be informed through consultation with relevant internal or external expertise and should be approved at the most senior level of the firm. The plan should be made public and circulated both internally to personnel in the firm and externally to partners and other relevant parties. Company policies and procedures should be coherent with the plan and provide accountability and incentives for aligning business activities with the stated commitments. The precise issues covered by the policy statement will depend on the actor's position in the value chain. More detailed

recommendations for what each set of actors should cover in these statements is provided in the following sections, but in general all actors can include commitments such as:

- Avoiding bias in training data sets and algorithms and practicing transparency regarding the same, either in development or purchasing decisions;
- Screening customers, geographies, and particular intended use cases to evaluate human rights risks;
- Avoiding the use of a company's products or services in certain types of deployments due to human rights concerns;
- Deploying internal processes and management structures to assess human rights risks on an ongoing basis and escalate concerns as needed;
- Supporting transparency and engagement, such as by including external advisers when making decisions involving FRT or by publicly reporting on customers, deployment, and efforts to avoid non-discrimination; and
- Requiring notification and consent and enabling remedy, as applicable.

Such policies should reflect the potential impacts of FRT on the full spectrum of human rights, not only privacy and non-discrimination.

Companies should undertake regular, independent audits to assess their adherence to these policies and practices. The results of these audits should be made publicly available to improve public trust and ensure that any issues that emerge are addressed in a timely manner.

Assessment of Impacts: Every actor in the FRT value chain should assess the actual and potential human rights impacts of its products and services. This should consider not only the actor's own impacts but also how it could contribute or be directly linked to the impacts of others. Assessments should focus on identifying who will be impacted by the technology, what the actual and potential impacts could be in both the near and long term, and what mechanisms exist to allow the company to create safeguards.

Typically, companies start this process with a high-level assessment of potential impacts. They then become more granular in their analyses, for example, examining issues such as bias in algorithms and training sets, issues related to particular product use cases, the rule of law in countries of deployment, and the reputations of customers and how they will use the technology. The assessment process should include technical, legal, and human rights experts from within the company, as well as outside stakeholder groups and representatives from the communities that may be impacted by the product.

Integration and Action Upon Findings: The precise way that any company integrates human rights impact findings into its decisionmaking will differ. Because the risks associated with FRT for many actors in the value chain result from customer use, senior-level oversight is necessary to overcome the urge to make sales regardless of the consequences. Relatedly, each actor will need to evaluate how to align performance incentives so that the identification of human rights risks and decisions not to sell to problematic customers are considered positively during performance reviews and in compensation.

Tracking the Effectiveness of Responses: The ability of companies to track the effectiveness of their efforts to address or mitigate impacts will vary, particularly when those impacts result from the actions of customers. Companies should put into place mechanisms to help them do so, including contractual mechanisms.

Communicating How Impacts Are Addressed: Greater transparency is urgently needed across the value chain. Companies at all stages of the FRT value development process should identify how they are managing potential and actual human rights impacts through public reporting. The reporting should focus on the most salient risks and potential on rights holders at the company's stage of the value chain. Increased transparency on outcomes and use cases would help companies in this sector come out of the shadows and differentiate themselves from less responsible competitors. It could help alleviate concern that the technology is being used in unknown ways and provide greater confidence that issues such as bias have been addressed. Greater transparency with customers and the public regarding the use of FRT in contexts where it is being deployed is also critical.

Remedy: So that grievances can be addressed early and remediated directly, business enterprises should establish or participate in effective operational-level grievance mechanisms for individuals and communities that may be adversely affected by impacts that the company caused or to which it contributed.[29] Companies are encouraged to support remedy for impacts to which they are directly linked. The line between when companies contribute instead of being directly linked is often difficult to distinguish and is not the focus of this paper. It is worth noting, however, that multiple actors in the value chain could be responsible for remedy.[30] In some instances, developers may be seen to have contributed to impacts if they do not put in adequate safeguards regarding their customers and, therefore, might also be expected to provide remedy. Other actors in the supply chain should, at a minimum, consider how product design and practices around transparency, notice, and consent can support remedy. This is an area that is ripe for further examination. Of course, the state also has a duty to ensure remedy is available.

Recommendations for Training Data Collectors

1. **Proactively share information about the source, demographic composition, and other details of training data sets with algorithm developers.**

   A significant portion of the demographic bias that has been observed in facial recognition systems can be attributed to the lack of representation of certain groups in the data used to train algorithms. To better allow facial recognition developers to assess the quality of a data set and the risks of training a model on any given data set, the groups compiling training data should measure the demographic composition of their images and make that information, along with other details about the provenance and context of the data set, available to developers who use it. Ideally, this information should also be made publicly available, and there should be opportunities for independent, third-party review of training data set composition.

2. **Assess privacy implications prior to the assembly of a training data product.**

   Prior to assembling a training data set, firms should conduct an assessment to identify potential risks from the collection, aggregation, and use of that data. Firms should consider the privacy impacts on the data subjects by considering the context in which those individuals shared their photos, their expectations as to how the images would be used, whether they had given consent for their data's use in the context of algorithmic training, and what the likelihood and consequences of a data breach may be. Firms should ensure that these assessments guide subsequent policies and practices and make arrangements for continual review of rights impacts.

3. **Establish policies of data security and collection minimization to reduce the risks of unauthorized access.**

   Training data managers should institute policies to protect training data from unauthorized access, including encryption, access controls, employee training, antivirus software, and other standard security practices. Training data collectors should not link biographic information about training data subjects to biometric data. This will reduce privacy risks in the event of unauthorized access. If biometric information must be collected, it should be stored and encrypted separately where possible.

4. **Establish systems to allow individuals to determine whether they are included as part of a training data set and to request removal when applicable.**

   Entities that collect images for training data sets should allow members of the public to submit requests to know whether they are included in the data set and to request removal. This process should be simple and easily accessible, and have a transparent process for follow-up. Firms should not begin or expand the collection and linking of biographic information to biometric data in order to enable these policies. Companies should consider only using training data sets where explicit consent has been obtained whenever possible, so long as they are still able to collect sufficiently representative data to reduce risks of bias.

5. **Conduct due diligence on potential buyers of training data to assess the risk of models leading to human rights abuses.**

   Before selling or sharing training data sets with algorithm developers, training data collectors should conduct an assessment of the potential recipient to determine whether that company, or the operators it sells to, is likely to be involved in any activities that could create human rights risks.

6. **Develop contractual clauses that limit the use of the training data and allow the collector to sever ties if evidence of abuse emerges.**

   Training data collectors who provide their data to external developers can include human rights safeguards language in contracts to limit the use of their training data and the products built from it to certain, agreed purposes and exclude particularly problematic uses. The collector should also require physical, technical, and organizational security measures to be implemented to protect the data from unauthorized access. The collector can require developers to include human rights safeguards language in end-user license agreements and require them to conduct their own human rights due diligence before selling products built on their training data. The ability of training data collectors to know that such restrictions were followed and act punitively if they were not might be limited, although they could refuse future sales to an entity known to have breached the contract.

7. **Be transparent about company policies and practices relating to data collection.**

   Training data collectors should use appropriate mechanisms—including their policy statement outlining their human rights commitments—to publicly share information about the company's policies and practices. This can include providing information about the sources of

data they use to compile training data sets, their policies surrounding subject consent, what mechanisms they have made available to allow individuals to query whether they are included in training data sets, and how they evaluate their customers. Training data collectors should commit to providing full and complete information about the provenance and demographic makeup of training data sets to potential buyers.

Recommendations for Algorithm Developers

1. **Evaluate any training data sourced from outside providers to ensure it is demographically representative of the population likely to be affected by the algorithm in development.**

   The use of unrepresentative data sets is one of the leading causes of demographic bias in facial recognition accuracy. Developers who source their training data from other groups must gain clear details and assurances from that group of the provenance and representation of the data set, or else develop a process for evaluating the demographics of the data themselves. While it is not possible to tell from the demographic makeup of the training data alone whether the resulting algorithm will be free from demographic biases, it is difficult to build accurate and unbiased systems from data that is not representative of the populations they will be used on. Developers who source their own training data should adhere to the recommendations included in the previous section.

2. **Rigorously test the performance of facial recognition models to determine their accuracy and to identify whether any demographic biases are present.**

   Algorithm developers should design and perform technical accuracy assessments designed to emulate the anticipated context of deployment as closely as possible. These tests should involve images for which the pose, lighting, camera angles, subject awareness, and overall quality is roughly equivalent to what will be encountered during deployment. If the algorithm is designed for multiple use contexts, it should be evaluated under multiple regimes. If possible, developers should test their algorithms in a simulated scenario environment using volunteers to gather better data about the operational accuracy of the deployed system.

   This process must also incorporate testing to determine whether the algorithm exhibits significant differences in accuracy for different demographic groups. The developer should determine an acceptable threshold for demographic differentials and refuse to ship any model that does not meet this standard. Some companies, such as Amazon, have dedicated fairness by design teams who implement these policies by working with new products and features to meet certain company standards prior to launch. Developers without these teams should consider implementing new organizational structures to ensure that a full range of potential impacts are taken into consideration when designing and conducting tests.

   To the extent possible, developers should make APIs and software toolchains available so independent researchers may conduct audits of the algorithm or otherwise design some process for working with third parties interested in conducting algorithmic assessments. Developers should also submit their algorithms to major third-party assessments, such as NIST's FRVT, to help contribute to development of common standards for the industry and aid operators in judging the relative technical capabilities of the vendors they are considering.

Information gathered through this testing process—including not only the results but the benchmarking data sets used, general limitations of the system, and common failure cases—should be made available to potential customers, and ideally be made publicly available. Developers should consider adopting standardized forms of communicating the details of their models, such as Google's Model Cards proposal.[31]

3. **Conduct due diligence on any software integrators that models are licensed to.**

   Before licensing a facial recognition model to be used as part of another company's software suite, facial recognition developers should conduct an assessment of the potential buyer to determine whether that company, or the operators it sells to, is likely to be involved in any activities that could create human rights risks. This requires considering the proposed activities of the customer, the customer's human rights reputation and past impacts, the products that the customer is known to have provided in the past, and any risks if the FRT model were to be used beyond the agreed-upon scope.

4. **Develop contractual clauses that limit the use of the algorithm and allow the developer to sever ties if evidence of abuse emerges.**

   Developers should insert human rights safeguards language into their contracts and licensing agreements that establish clear use limitations and which would allow them to withdraw access to the model, or limit access to model updates and support, if the seller is found to be using it to engage in human rights abuses or is supplying operators who engage in human rights abuses. The developer can require integrators to include human rights safeguards language in end-user license agreements and require them to conduct their own human rights due diligence before selling products built from their models. Companies could also require their customers to publicly state that their algorithm is being used, which could assist with remedy.

5. **Practice transparency about company policies and practices related to algorithm development.**

   Algorithm developers should use appropriate mechanisms—including their policy statement outlining their human rights commitments—to publicly share information about the company's policies and practices. This should include not only reporting the results of any performance evaluations but also information about the data sets used for these tests, the conditions under which they occurred, and whether any difference was measured between demographic groups. Developers should strive to ensure their testing and reporting conforms to international standards and industry best practices. Developers should also proactively communicate the policies, limitations, and safeguards they have in place regarding selling to certain types of customers or for certain use cases.

Recommendations for Software Integrators

1. **Perform rigorous testing on models purchased from outside developers to determine their accuracy and ensure they are free from demographic biases.**

   When software integrators source their algorithms from external firms, they should have a process in place to evaluate potential models for their accuracy and bias. Firms should prioritize developers that are able to provide independent assessments proving the model's

accuracy and documentation attributing to the fact that the models had been trained on diverse data sets that include members of the groups likely to encounter the system during deployment. Information regarding the testing, performance, and limitations of facial recognition algorithms should be made available to all of the integrator's potential customers to help them assess whether and how to use the products.

2. **Institute internal structures and processes for identifying and escalating the potential human rights concerns posed by new products and services.**

   When developing a new facial recognition product or service, firms should go through a formal process to identify the potential risks before it is released. Firms should consider, for example, the privacy risks from having sensitive information collected and stored about a subject, the possibility that operators may use these capabilities to discriminate against certain groups, and the potential harm of errors made by the system. Firms should establish safeguards to manage the risks identified through this assessment process and ensure that the assessment is regularly updated to account for changes in the technology and in the context of its use. The depth of the assessment may vary depending on the severity of the initial findings.

   Integrators should institute internal mechanisms for reviewing the development process for new products and services. These structures should allow and encourage employees to voice concerns that come up during their work and escalate those concerns as necessary to a specialized body with the authority to set company policy. This body should be made up of a diverse group, including not only the company's legal team but also representatives from product development and customer support.

   One example of a facial recognition developer with these kinds of institutional safeguards is Microsoft. Microsoft's AI, Ethics, and Effects in Engineering and Research (AETHER) Committee provides the company with an internal structure for dealing with cases where the potential risks of a new product or use case is deemed too dangerous. This process has already led Microsoft to refuse potential contracts that would, for example, allow their facial recognition systems to be used for real-time identification in police body cameras.

3. **Establish external advisory bodies with representatives from a wide range of disciplines to provide outside assessment of the propriety of potentially risky new products or services.**

   Integrators should consider establishing external advisory bodies to provide independent accountability and advice on issues relating to whether and how to bring new facial recognition products and services to market. These advisory bodies should be composed of experts from a range of disciplines, including human rights, consumer protection, accessibility, data science, and data protection. Firms should develop a structured process for bringing issues to this body for consultation, and all deliberations should be made publicly accessible, possibly with a time lag to manage any commercial sensitivities.

4. **Practice principles of privacy and data protection by design and default.**

   Integrators should incorporate technical controls into the design and architecture of facial recognition systems to enforce, throughout the full life cycle of the data being collected, privacy and data protection principles, including transparency, security, integrity, access

control, accountability, and minimization. Software developers should proactively consider the potential for privacy violations arising from errors or intentional misuse and design safeguards to guard against these risks. Examples of relevant practices could include designing systems to:

- Ensure that data is encrypted and stored securely in a way that only the operator can access;
- Segregate biometric data from other personal data;
- Automatically delete any information collected from individuals who are not matched by the system;
- Only store facial templates rather than raw images;
- Regularly delete or prompt operators to review records that have been retained for a certain amount of time;
- Only store the minimum necessary metadata to match records; and
- Have appropriate default confidence thresholds for searches.

Integrators should implement organizational practices to ensure privacy by design is followed and enforced, such as conducting regular internal reviews, assigning dedicated personnel to oversee privacy issues, and training employees on privacy.

5. **Conduct due diligence of potential buyers to assess the risk of deployments leading to human rights abuses.**

Before providing facial recognition platforms or services to an operator, integrators should conduct due diligence on the potential buyer to determine whether the organization is likely to be involved in any activities that could create human rights risks. Technology developers should develop a set of principles regarding what they consider a responsible use of their technology and the conditions that would serve as red lines for what they consider acceptable use. These principles should be made publicly available and be incorporated into an internal process of reviewing potential new customers to determine whether a sale will lead to risk of abuse.

Assessments of potential customers should be based on a number of factors, including the legal environment of the country where the organization is located, the system's intended scope and purpose of use, the past activities and rights impacts of the operator, and the safeguards and governance procedures that the organization has put into place. For a more detailed discussion of due diligence considerations and red flags, companies can reference the U.S. Department of State Guidance on Implementing the UNGPs.[32]

Microsoft is an example of a company that has put such a process in place, through its AETHER Committee, which has already led to the company rejecting at least one potential government customer due to human rights concerns. The success of this process demonstrates the importance of conducting due diligence at an early stage of the contracting process and ensuring senior-level involvement in vetting decisions.

For firms that provide facial recognition as a service to many small operators via the cloud, safeguards should include, at a minimum, Know-Your-Customer requirements for those who sign up to use the service and a process for identifying users whose activity suggests that they may be at high risk of contributing to rights abuses.

6. **Leverage contractual or other mechanisms to establish processes for controlling or regularly reviewing how customers are using the tools being provided to them.**

   Integrators should insert terms into their contracts and licensing agreements that prohibit operators from using their products or services in ways that could violate the rights of others. These terms could also require that the customer provide notice and require consent from those included in matching data sets. This would help enable remedy where applicable. Oversight and enforcement of these terms can be accomplished through regular audits by the developer—which could be tied to licensing sunsets—or through observation of the operator's deployment by customer support, professional service, and sales teams as part of an ongoing relationship between the organizations.

   Developers offering facial recognition as a cloud service should compensate for their relative lack of oversight opportunities by implementing other measures, such as limiting the volume of images that can be processed or the number of licenses that can be in use, in order to protect against abuses. Facial recognition providers should cut off access to their services, refuse to renew licenses, or deny access to software updates and technical support for operators that are found to be in breach of these terms and who fail to rectify the issues once they have been raised. However, these options may not be readily available or trackable for cloud products, making it all the more important that human rights are considered in their initial design. To the extent possible, all facial recognition integrators should strive to implement technical mechanisms for gathering data that could assist in carrying out ongoing oversight into how their products and services are being used by customers.

   Another way developers can retain control over the use of their technologies is by allowing some features to be sold directly to end users to deploy and operate independently while requiring high-risk uses be filtered through the developer. One facial recognition provider interviewed for this report had implemented such a system for one of their law enforcement clients. The developer had helped the client install the capacity to independently operate live monitoring at certain sites for limited periods of time during specific high-profile events but required them to submit requests for retroactive identifications to the developer. This creates an additional layer of accountability that can help prevent abuses. In general, live monitoring should be considered the higher priority for additional oversight, but integrators should take into account the potential scope and impact of specific deployments when determining what appropriate accountability measures should look like.

   Firms may encounter difficulties in finding out when contract provisions have been breached and enforcing the terms of their agreements. This should not, however, be taken as a reason for not using every mechanism available to avoid the abuse and misuse of this technology, given the magnitude of potential harms.

7. **Provide rigorous and accessible training for customers to help operators understand how to use the technology in ways that respect human rights.**

   Prior to use, technology developers should require customers to undergo training on the capabilities and limitations of facial recognition systems as well as proper procedures to help ensure responsible use. At a minimum, this training should include information on how to

evaluate matches returned by the system, how to adjust confidence intervals for different use cases, what the sources and rates of error are during real-world deployments, and how to ensure that data collection is reduced to the minimum necessary for operation. Developers should also provide guidance to operators on how to provide notice to data subjects in situations or areas where individuals will be subject to facial recognition, and generally encourage transparency in the operations of their buyers.

8. **Be transparent about company policies and practices relating to software integration and sale.**

   Software integrators should use appropriate mechanisms—including their policy statement outlining their human rights commitments—to publicly share information about the company's policies and practices. This should include how they evaluate the algorithms sourced for their products and services, what purposes they will or will not allow their products to be used for, how they assess potential customers for human rights risks, how their development process incorporates the principles of privacy by design, whether and how they exercise oversight into how their products are being used after sale, and how they train operators to use their systems responsibly.

6

Collective Action

Industry Approaches and Regulation

This analysis has focused primarily on what companies should do individually to respect human rights when developing and selling FRT systems. However, there is a need for collective action in some areas, and governments also can play a critical role in ensuring that private sector actors throughout the FRT value chain have in place appropriate human rights safeguards.

Improving Accuracy and Eliminating Bias

Better methods of communicating the composition of training data sets and the performance of algorithms through the FRT value chain would help award good actors and ensure that an understanding of the risk of inaccuracy and bias carries through all the way to the operator. Developers should work together to encourage the adoption of common methods of testing and reporting for facial recognition systems, such as the new ISO/IEC 19795-1 standard, and develop industry-wide practices for the measurement and reporting of details about training data sets. FRT providers should also collaborate in establishing certifications schemes attesting to the performance of facial recognition products and services, building on the progress of groups such as the FIDO Alliance and recent work by academic and government researchers. Such certification or ranking should ideally include not only technical testing of the underlying algorithms but also simulated testing of the full system to better assess the predicted real-world performance of the service.

Ideally, governments would be able to require that facial recognition software meet certain standards of accuracy and non-discrimination and be certified to them before deployment. In reality, however, it is currently difficult to settle on common standards of performance given the wide diversity of contexts FRT can be used for, the multiple different metrics FRT performance can be measured

along, and the challenge of predicting real-world performance based on data gathered in a controlled laboratory setting. However, ongoing work by industry and by government and academic researchers can help make progress toward this goal.

Governments should remain engaged in this work, with the goal of eventually devising frameworks for certifying FRT performance, beginning with high-risk uses such as law enforcement deployments. Governments will then need to identify bodies that are trusted to evaluate systems according to these standards. Eventually, governments may be able to require certification for any system being used by public agencies in order to assure the industry that there will be a reliable market for certified software. These standards can then serve as a quality signal for private sector operators, further incentivizing FRT developers to ensure their products and services meet these benchmarks.

Even in the absence of formal certification schemes, governments should, at a minimum, require that any agency using FRT collects and publicly reports details about the algorithm's performance (including any demographic effects) and the provenance of training data involved in the systems they use. They can also require developers to make their algorithms available for third-party testing. This will limit government agencies to sourcing from developers that practice transparency in their development process, helping to push the industry toward common standards of testing and reporting.

Privacy, Notice, and Consent

The vast majority of training data used for modern facial recognition systems is collected without the explicit consent of the subjects. While it may not be possible to change this, given the current state of the industry, general data privacy and data protection laws can help provide clarity as to which collection strategies may be allowed and which strategies—such as data sharing between firms or collection from operational systems—may require additional notice or consent requirements to be lawful.

Laws can also establish the rights of individuals to query operators as to whether they are included in existing databases and to request that their information be removed in a timely manner. To protect this information from unauthorized access, policymakers should also establish minimum requirements for data security and set expectations around collection, minimization, and the enforcement of retention limits for any data associated with facial recognition development.

Human Rights Due Diligence

Governments can lay out broad frameworks that companies in the FRT value chain should follow regarding human rights due diligence processes. This could help engage new sectors deploying the technology and also prompt more responsible behavior by smaller actors in the value chain that may not be part of ongoing discussions on FRT and human rights and ethics.

A key part of this should be improving communication between developers and the teams in government that are aware of the specific risk profiles of governments and corporations around the world. Interviews for this report indicate that developers often feel that they could make better decisions about who to sell to if they had more information from government about the legal and human rights environments of countries around the world and what abuses have been connected to specific organizations in the past. These decisions would also be aided by governments providing

clear direction about what they consider to be impermissible uses of FRT. If certain red lines could be established around use cases, such as those involving mass surveillance or the collection of biometric information from children, facial recognition providers could use those restrictions as a starting point for their own policies about who they will agree to sell to.

Remedy

Governments should establish penalties for human rights abuses arising from FRT. Transparency about the use of FRT, as outlined above, is a critical precursor for remedy, given that the technology can be deployed without peoples' knowledge.

Enforcement will likely be multifaceted. Individuals and groups should be able to bring cases when their rights are impacted by FRT. Given the potentially large groups affected by issues such as bias, class actions must be permitted.

Regulators with adequate technical backgrounds should be tasked with proactively and independently carrying out risk-based audits of companies in the FRT value chain, focusing particularly on issues of non-discrimination and adherence to requirements regarding privacy, data-sharing, and notice and consent.

About the Authors

Amy K. Lehr is a senior associate with the Human Rights Initiative at the Center for Strategic and International Studies (CSIS). She is the former director of the CSIS Human Rights Initiative and was a senior fellow with the program. In that role, her work focused on human rights as a core element of U.S. leadership, labor rights, emerging technologies, and the nexus of human rights and conflict. Amy previously served as legal adviser to the UN special representative on business and human rights and helped develop the UN Guiding Principles on Business and Human Rights. She was a fellow at the Harvard Kennedy School's Corporate Responsibility Initiative. Amy formed part of a business and human rights legal practice, engaging with businesses, investors, multilateral organizations, civil society, and governments to address global human rights challenges. She previously worked for development nongovernmental organizations in Myanmar and Thailand. She was a Council on Foreign Relations term member. Amy received her AB from Princeton and her JD from Harvard Law School.

William Crumpler is a research associate with the Strategic Technologies Program at CSIS, where his work focuses on cybersecurity policy and the governance of artificial intelligence and other emerging technologies. He holds a BS in materials science and engineering from North Carolina State University.

Endnotes

1 Inioluwa Deborah Raji and Genevieve Fried, "About Face: A Survey of Facial Recognition Evaluation," Association for the Advancement of Artificial Intelligence, 2021, https://arxiv.org/pdf/2102.00813.pdf.

2 Olivia Solon and Cyrus Farivar, "Millions of people uploaded photos to the Ever app. Then the company used them to develop facial recognition tools.," NBC News, May 9, 2019, https://www.nbcnews.com/tech/security/millions-people-uploaded-photos-ever-app-then-company-used-them-n1003371.

3 Cade Metz, "Facial Recognition Tech Is Growing Stronger, Thanks to Your Face," *New York Times*, July 13, 2019, https://www.nytimes.com/2019/07/13/technology/databases-faces-facial-recognition-technology.html.

4 Madhumita Murgia, "Who's using your face? The ugly truth about facial recognition," *Financial Times*, September 18, 2019, https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e; Olivia Solon, "Facial recognition's 'dirty little secret': Millions of online photos scraped without consent," NBC News, March 12, 2019, https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921; and Madhumita Murgia, "Microsoft quietly deletes largest public face recognition data set," *Financial Times*, June 6, 2019, https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2.

5 Harvey Adam and Jules LaPlace, "Exposing.ai," Exposing.ai, https://exposing.ai; and Metz, "Facial Recognition Tech Is Growing Stronger, Thanks to Your Face."

6 John R. Smith, "IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems," IBM Research Blog, January 29, 2019, https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/.

7 "MORPH Facial Recognition Database," University of North Carolina Wilmington, n.d., https://uncw.edu/oic/tech/morph.html; Murgia, "Who's using your face? The ugly truth about facial recognition"; and Metz, Facial Recognition Tech Is Growing Stronger, Thanks to Your Face."

8 Kyle Wiggers, "Facebook dataset combats AI bias by having people self-identify age and gender," VentureBeat, April 8, 2021, https://venturebeat.com/2021/04/08/facebook-dataset-combats-ai-bias-by-having-people-self-identify-age-and-gender/.

9 "Training Data for Facial Recognition," Lionbridge Technologies, n.d., https://lionbridge.ai/solutions/facial-recognition-data/; and "Face recognition training data: helping to train a software," Clickworker, n.d., https://www.clickworker.com/case-studies/training-data-for-a-face-recognition-software/.

10 Amy Hawkins, "Beijing's Big Brother Tech Needs African Faces," *Foreign Policy*, July 24, 2018, https://foreignpolicy.com/2018/07/24/beijings-big-brother-tech-needs-african-faces/.

11 William Crumpler and James A. Lewis, *How Does Facial Recognition Work? A Primer* (Washington, DC, CSIS, 2021), https://www.csis.org/analysis/how-does-facial-recognition-work.

12 Sandra Azria and Frédéric Wickert, *Facial Recognition: Current Situation and Challenges* (Strasbourg, France: Council of Europe, 2019), https://rm.coe.int/t-pd-2019-05rev-facial-recognition-report-003-/16809eadf1.

13 Patrick Grother, Mei Ngan, and Kayee Hanaoka, "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects," National Institute of Standards and Technology (NIST), 2019, doi:10.6028/NIST.IR.8280.

14 Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker, "Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings," ECCV, 2018, https://arxiv.org/abs/1809.02169; Alexander Amini et al., "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure," Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), https://lmrt.mit.edu/sites/default/files/AIES-19_paper_220.pdf; Boyu Lu et al., "An Experimental Evaluation of Covariates Effects on Unconstrained Face Verification," *Journal of Latex Class Files* 18, no. 4 (2018), https://arxiv.org/pdf/1808.05508.pdf; and Martins Bruveris et al., "Reducing Geographic Performance Differentials for Face Recognition," (WACVW, 2020): 98-106, https://ieeexplore.ieee.org/document/9096930.

15 "Face Recognition with OpenCV," OpenCV, n.d., https://docs.opencv.org/2.4/modules/contrib/doc/facerec/facerec_tutorial.html; "Open Source Biometric Recognition," OpenBR, n.d., http://openbiometrics.org/; and Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan, "OpenFace," CMU School of Computer Science, 2016, https://cmusatyalab.github.io/openface/#openface.

16 Nicolás Rivero, "The little-known AI firms whose facial recognition tech led to a false arrest," Quartz, June 26, 2020, https://qz.com/1873731/the-unknown-firms-whose-facial-recognition-led-to-a-false-arrest/.

17 Brian Rhodes, *Alcatraz Presents Face Recognition Access Control* (Bethlehem, PA: IPVM, June 2020), https://ipvm.com/reports/alcatraz-presents.

18 John Honovich, *Foolish Strategy: OEMing Facial Recognition* (Bethlehem, PA: IPVM, December 2018), https://ipvm.com/reports/oem-face; and Charles Rollet, *19 Facial Recognition Providers Profiled* (Bethlehem, PA: IPVM, April 2019), https://ipvm.com/reports/face-iscw-19.

19 Bethan Davies, Martin Innes, and Andrew Dawson, *An Evaluation of South Wales Police's Use of Automated Facial Recognition* (Cardiff University, 2018), https://afr.south-wales.police.uk/wp-content/uploads/2019/10/AFR-EVALUATION-REPORT-FINAL-SEPTEMBER-2018.pdf; and Curtis Waltman, "California Department of Justice spent nearly two million dollars on controversial facial recognition software," MuckRock, April 27, 2017, https://www.muckrock.com/news/archives/2017/apr/27/california-doj-facial-recognition/.

20 Chris Dulhanty and Alexander Wong, "Investigating the Impact of Inclusion in Face Recognition Training Data on Individual Face Identification," AAAI/ACM Conference on AI, Ethics, and Society (AIES 20), 2020, https://arxiv.org/abs/2001.03071.

21 Patrick Grother, Mei Ngan, and Kayee Hanaoka, "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects."

22 Hu Han and Anil K. Jain, *Age, Gender and Race Estimation from Unconstrained Face Images* (East Lansing, Michigan: Michigan State University, 2014), http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf.

23 Patrick Grother, Mei Ngan, and Kayee Hanaoka, "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects."

24 Alvi, Zisserman, and Nellaker, "Turning a Blind Eye"; Amini et al., "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure"; Boyu Lu et al., "An Experimental Evaluation of Covariates Effects on Unconstrained Face Verification"; Martins Bruveris et al., "Reducing Geographic Performance Differentials for Face Recognition."

25 Inioluwa Deborah Raji and Joy Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," Association for the Advancement of Artificial Intelligence, 2019, https://dam-prod.media.mit.edu/x/2019/01/24/AIES-19_paper_223.pdf.

26 Drew Harwell and Eva Dou, "Huawei tested AI software that could recognize Uighur minorities and alert police, report says," *Washington Post*, December 8, 2020, https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/; "Alibaba facial recognition tech specifically picks out Uighur minority – report," Reuters, December 16, 2020, https://www.reuters.com/article/us-alibaba-surveillance/alibaba-facial-recognition-tech-specifically-picks-out-uighur-minority-report-idUSKBN28R0IR; and Johana Bhuiyan, "Major camera company can sort people by race, alert police when it spots Uighurs," *Los Angeles Times*, February 9, 2021, https://www.latimes.com/business/technology/story/2021-02-09/dahua-facial-recognition-china-surveillance-uighur.

27 United Nations Human Rights Office of the High Commissioner, *Guiding Principles on Business and Human Rights* (New York: United Nations, 2011), https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf.

28 "B-Tech Project," United Nations Human Rights Office of the High Commissioner, https://www.ohchr.org/EN/Issues/Business/Pages/B-TechProject.aspx.

29 United Nations Human Rights Office of the High Commissioner, *Guiding Principles on Business and Human Rights*.

30 United Nations Human Rights Office of the High Commissioner, "Access to remedy and the technology sector: basic concepts and principles," A B-Tech Foundational Paper, January, 2021, https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf.

31 Huanming Fang and Hui Miao, "Introducing the Model Card Toolkit for Easier Model Transparency Reporting," Google AI Blog, July 29, 2020, https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html.

32 "Guidance on Implementing the UN Guiding Principles for Transactions Linked to Foreign Government End-Users for Products or Services with Surveillance Capabilities," U.S. Department of State, 2020, https://www.state.gov/wp-content/uploads/2020/10/DRL-Industry-Guidance-Project-FINAL-1-pager-508-1.pdf.

COVER PHOTO ADOBE STOCK

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

1616 Rhode Island Avenue NW
Washington, DC 20036
202 887 0200 | **www.csis.org**