WHAT'S IN YOUR FACE? DISCRIMINATION IN FACIAL RECOGNITION TECHNOLOGY

A Thesis
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Masters of Arts
In Communication, Culture, and Technology

By

Jieshu Wang, M. Eng.

Washington, DC
April 13, 2018

WHAT'S IN YOUR FACE? DISCRIMINATION IN FACIAL RECOGNITION TECHNOLOGY

Jieshu Wang, M. Eng.

Thesis Advisor: Mark MacCarthy, Ph.D.

ABSTRACT

This paper examines the discrimination in facial recognition technology (FRT) and how to mitigate it in the contexts of academia, product development, and industrial research. FRT is the automation of the processing of human faces. In recent years, given the fast development of machine learning techniques, FRT gained considerable momentum. FRT is increasingly trained on extraordinarily large datasets and sophisticated algorithms, and its accuracy has been increased to the point that surpasses human capacity. Applications of FRT emerge in a variety of fields, such as surveillance, military, security, and e-commerce. At the same time, many ethical issues have been raised. In this paper, two types of FRT applications are distinguished—identification and classification. The former aims to search and match the captured face in the target database to pinpoint the identity, while the latter classifies people into different groups according to some properties drawn from their facial features, for example, gender, race, age, and sexual orientation. The latter type raises serious discrimination issues, because the training data is inherently biased, and it could be easily used to develop discriminatory applications and increase the number of people who suffer from discrimination. In order to mitigate the discrimination issue, three types of FRT design practices are identified—product development, academic research,

and industrial research. Value Sensitive Design (VSD) is a helpful approach to minimize discriminatory issues in product development. In academic settings, the traditional way to ensure ethical outcomes is through institutional review boards (IRB), but IRB has many disadvantages when dealing with FRT and data science in general. In industrial research, Facebook's ethical review system developed after the "emotion contagion" study is discussed as a case study to demonstrate general principles that could help private companies in the FRT field to mitigate discrimination issues in research, such as ethical training and building multidisciplinary reviewing teams.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1. Controversies

One's face is among the most reliable and public aspects of human's identity and personality. The unique combinations of facial features show who we are, and the expressions on faces reveal our feelings, emotions, and even underlying intentions. Since the debut of the "Computer Physiognomy" of Nippon Electric Company (NEC) at the World's Fair in Japan in 1970, computer scientists and engineers have devoted considerable efforts attempting to automate the processes of recognizing people facially and analyzing the related information using machines. This endeavor is named Facial Recognition Technology (FRT). Since then, dramatic improvements have been made in the field of FRT, along with the development of digital cameras, storage hardware, high-speed network, and most notably, computing techniques such as Artificial Intelligence (AI). Today, the accuracy of FRT is claimed to be more advanced, and it is favored in a wide range of applications, both governmental and commercial.

Despite the enormous practical potential, growing concerns are raised with each step forward regarding ethical, legal, and policy-making issues. For example, in September 2017, a Stanford study featured in *the Economist* that claimed to use FRT to identify gay people[1] drew much attention and sparked fierce objections among LGBTQ groups. The researchers created a sexual-orientation classifier using deep neural networks trained on a database containing 35,326 facial images. According to their research paper, the classifier could correctly distinguish homosexual people with an accuracy approximately 20% higher

than humans and the result was consistent with the facial features predicted by the prenatal hormone theory of sexual orientation.[2] Two of the most prominent LGBTQ organization in the US criticized the study as "dangerous and flawed" and with the potential to put homosexual people at risk. However, Michal Kosinski, co-author of the paper, interestingly noted that he considered this research support for LGBTQ rights because it provided evidence to the biological basis of sexual orientation.[3]

Another recent controversial event is a criminal-face-prediction study conducted by the researchers at Shanghai Jiaotong University in China in 2016.[4] In the study, a convolutional neural network was trained on 1,856 ID photos of Chinese men between 18 and 55 years old, half of whom were convicted criminals, and it was claimed to be able to distinguish criminals and non-criminals with an 89.5% accuracy. Although the researchers admitted that much more work needs to be done before a strong statement can be made, a heated debate was caused on the consequences of using this kind of FRT to identify potential criminals even before they commit a crime, like the scenario in the *Minority Report*, or to score people based on their possibility to break the law given their facial appearance, like the scenario in the *Black Mirror*.[5] Although guidelines are formed to ensure that academic research aligns with human value, such as Belmont Report, no specific consensus has been reached on whether this kind of FRT research should be carried out and published.

Not only in academia but also in industry and everyday life FRT gives rise to controversies. The first major debate on privacy broke out in June 2001 when FRT was found to be used on thousands of football fans in the Super Bowl in Tampa, Florida, snapping every spectator and matching their faces against a database of criminals. Since

then, this Super Bowl has been called "Snooper Bowl" ironically.[6] Concerns about privacy skyrocketed. Norman Siegel, the former director of the New York chapter of the American Civil Liberties Union said the situation indicated that FRT was "outpacing the civil right and civil liberties along with the right of anonymity and privacy." At the same time, however, law enforcement departments consider FRT a "powerful tool to assist in maximizing public safety," as Detective Bill Todd Jr. of Tampa police department in Florida put it, refuting that FRT invaded people's privacy.[7] To people's surprise, however, besides Super Bowl, FRT had been used for various purposes all around the world for a long time, in some places even as a routine means of surveillance, such as London Borough of Newham, where a camera network equipped with FRT had been covering public areas since 1998.[8]

The tragedy of September 11 attack kindled the interests of U.S. governments in deploying FRT in public places such as airport for security and law enforcement purposes. Some experts from commercial FRT companies believe that FRT "could have instantly checked the image against photos of suspected terrorists on file with the FBI and other authorities,"[9] suggesting that September 11 attacks might have been avoided with the help of FRT. Now, some travelers departing from several major airports in the U.S. such as Boston's Logan International Airport have had their faces scanned and will soon be subject to FRT against Department of Homeland Security (DHS) databases.[10]

The big promise of government deployment has brought massive incentive to the research, development, and application of FRT in law enforcement and other scenarios such as identifying missing children. According to a study by Center on Privacy & Technology at Georgetown University, half American adults, that is, 117 million people,

are included in the law enforcement face recognition network, largely without knowing, and real-time FRT on surveillance cameras are being used today by major police departments, including Chicago and Los Angeles. Ohio's FRT system, which contains all Ohio state drivers' license photos, remained entirely unknown to the public for five years. The most significant problem, however, as identified in the study, is that the usage of FRT is unregulated, so that most law enforcement agencies using FRT do not consider many important issues, such as, free speech, accuracy, privacy, and bias.[11]

Recently, with the incorporation of AI techniques, FRT performance has skyrocketed, reaching or even surpassing human capacity without human intervention.[12] FRT can be used in a wide range of application scenarios, from airports to shopping malls, and to identity verification in payment transfer on cell phones, so it involves a variety of stakeholders and players, each of whom has her own, or even opposite moral stand, leading to conflict in opinions. For example, a police officer may think his moral responsibility is to ensure the safety of the community, while residents might feel their privacy is being intruded upon by the routine surveillance used to keep out terroristic threats. Also, FRT is a complex mixture of a large number of technologies, each of which involves different and intertwining ethical issues. For example, like computer technology, FRT involves many ethical issues discussed in computer and information ethics. Moreover, Having AI techniques inside, FRT touches on many issues involved in AI ethics, such as algorithmic bias.

One of the most prominent problems is the bias related to demographic information, including race, sex, and age. For example, it is known that the FRT used by the law enforcement agencies in the U.S. was trained on databases mostly populated with

information on white males, therefore, its accuracy in recognizing white males is considerably higher than the accuracy on females, children, and people of colors.[13] Similarly, MIT student Joy Buolamwini found her face unrecognizable to an FRT system when she visited Hong Kong because the algorithm was trained largely on Asian people, which prompted her to found the Algorithmic Justice League (AJL) to fight bias in machine learning.[14] This bias could lead to further problems. For example, in 2015, the FRT in Google Photo was found to tag two African Americans as "Gorillas," provoking trenchant criticism.[15] Also, the chance of an African-American woman being misidentified as a criminal in FBI's FRT system is much higher than that of a white male. Such misidentification can subject people of color to increased bureaucratic interference if FRTs with demographic biases are widely implemented throughout government agencies, wasting their time, energy, and money, indirectly increase inequality in opportunities.

2. Organization of the thesis

Two types of FRT applications

Regarding ethics, two types of applications of FRT can be distinguished—identification and classification. Identification is a search for a captured face image in a database in order to recognize the person's identity. Classification FRT is labelling of people according to their facial features into different groups without identifying who they are. Most FRT used for surveillance, security, law enforcement, and financial purposes use the identification type of FRT, for instance, the Super Bowl FRT in 2001. Identification FRT applications mainly capture faces through CCTV cameras, pull their facial features, and search for matches in a database of interests, such as a database of criminals, employees,

and clients, in order to identify or verify people's identities. Personally identifiable information (PII), or sensitive personal information (SPI) being involved, identification applications directly involve issues of privacy. People expect to go anywhere anonymously as long as they obey the laws. The deployment of FRT for identification purposes, however, greatly increase the likelihood that individuals will be tracked. Even if the faces are not identified immediately, the possibility of later identification always exists because the information is stored in a database. Helen Nissenbaum argued that justifiable privacy is still expected even in public areas, so public surveillance violates people's right to privacy because of its violation of "contextual integrity."[16] Also, Philip Brey identified an issue called "function creep," an expression he borrowed from John Woodward, concerning risks such as the possibility that FRT developed to identify criminals might extend for total surveillance and other unanticipated purposes through the widening of the database and the shifts of user and domain.[17]

On the other hand, classification FRT, the applications for labelling people according to their facial features involve a different set of ethical issues. The Stanford gay-recognition study and Shanghai Jiaotong University's criminal prediction study both belong to this type, which attempts to label or tag people, classifying them into different groups based on their facial appearances or expressions, instead of recognizing their identities simply because identities are irrelevant in these situations. There are many issues associated with this kind of research and practice. For example, the tags themselves are absolutely ambiguous. Many scholars like Alfred Kinsey believe that human sexual identity and orientation are continuously distributed on spectrums rather than simply falling into two or three categories.[18] Similarly, the standards to distinguish criminals and non-

criminals are largely based on in which country and in what historical period one is examining. In addition, these algorithms might not be accurate enough. For example, for the Stanford study, even if the biological base provides some clues for sexual orientation that are visible to machines, it fails to tell the whole picture. The face is just a very small part of the human body, therefore, the algorithm is unable to make a comprehensive prediction, not to mention its ignorance of the significant roles of life experience and psychological aspects in building sexual identity. Moreover, the tags in question might be very sensitive and private information that people don't want to disclose, thus, even if the algorithms are accurate enough, the application of these technologies might become an invasion of privacy. Most importantly, discrimination could easily occur in these applications. For example, if the Stanford algorithm is integrated into commercial scenarios, it could lead to discriminatory price, service, words, and behaviors based on sexual orientation. The criminal prediction study may cause discriminatory behaviors towards people with "criminal-inclined" faces even if they've never done anything illegal. Classifying people into different groups could also lead to discrimination and harassment based on sex, race, age, color, and other information in all kinds of social affairs such as employment, insurance, education, credit, medical care, and law enforcement, increasing inequality.

How the thesis is organized

Two types of FRT—identification and classification—cause different kinds of ethical issues. Identification FRT mainly violates privacy and autonomy, while classification FRT would cause discrimination issues. This paper focuses on the discriminatory issues of FRTs that classify people into different groups based on their facial

features. I'll discuss this issue through three lenses—technical, ethical, and design.

Chapter two introduces the history, theories, and techniques related to FRT. A brief technological history of FRT will be given. The early attempts of FRT require manually coding facial features into the algorithms. With the integration of machine learning, however, FRT gains significant momentum, being able to acquire high accuracy by training on large datasets of human faces. Today, the accuracy of FRT is even higher than human capacity, having huge potentials in many areas. The applications of FRT can be divided into two types—identification and classification. The former raises issues such as privacy, while the latter raises concerns about discrimination since the automation of social sorting could lead to discriminatory applications.

Chapter three focuses on the ethics of FRT-driven classification tasks. First, three ethical frameworks—utilitarianism, deontology, and virtue theory will be discussed. Then, I'll talk about the ethical issues of discrimination based on the three frameworks and why discrimination is morally wrong. I argue that FRTs with classification functions can easily raise discrimination issues.

Chapter four seeks to mitigate the discrimination issues of FRT through a design perspective. In other words, I argue that discrimination could be reduced if there's a way to incorporate the value of freedom-from-discrimination into the FRT practice. It is helpful to distinguish three types of FRT practices—commercial product development, academic research, and industrial research. For product development, I argue that using value sensitive design (VSD) approach could help reduce the ethical risks. VSD consists of three phases of investigations—conceptual, empirical, and technical. Suggestions in each phase are provided. For academic research, I argue that IRB has many limitations when

overseeing FRT research and data science in general because it was designed to protect human subjects in biomedical and behavioral research. The engagement of human subjects in FRT research is ambiguous and indirect, making it hard to pinpoint the potential harm. To adapt new situations in ICT research, the Menlo Report was released in 2012 by U.S. Department of Homeland Security, providing some new guidelines for FRT academic research. For industrial research, which is currently in a grey area, a case study on Facebook's ethical reviewing system will be discussed. Facebook's notorious "emotion contagion" study sparked widespread outrage, and Facebook quickly came up with an ethical review system for any future research. I argue that this system is a good example but is not universally helpful for FRT companies since not all companies have the resources to spend like Facebook.

CHAPTER 2

FACIAL RECOGNITION TECHNOLOGY IN A NUTSHELL

1. A brief history of Facial Recognition Technology

The idea of recognizing people from pictures roots in the very invention of cameras in the 1830s, before which the most efficient way to identify prisoners in England is branding. From 1852, as a more humane alternative that was dubbed "Angel copier," taking photos of prisoners became a routine in English prisons for the reasons of tracking down prison breakers and data sharing.[19] At the end of the 19th century, a French police officer named Alphonse Bertillon created a system that could identify people by measuring their facial features such as the shape of the ears and noses.[20]

The introduction of computers brought about the promise of automating the process of facial recognition. Due to the limited computing powers at that time, however, early FRT systems were only semi-automated, in need of manual coding of a lot of things in advance. In addition, face recognition was approached as a generic pattern recognition problem, with most methods were based on the geometric features of human faces. The first computer FRT system was created in the 1960s by Woodrow Bledsoe, a prominent mathematician and computer scientist, also one of the founders of AI. By manually locating "facial landmarks" with horizontal and vertical coordinates in pictures, the system could compute attributes associated with the landmarks, such as the width of mouth, the distance between eyes, the location of the hairline, and some ratios, and compare them against reference data in order to find a match in a database.[21] The accuracy of this kind of semi-automated FRT system was improved in the 1970s by Goldstein, Harmon, and Lesk, who added twenty-

two "subjectively judged 'features' descriptions" like long ears, lip thickness, and hair colors into the system.[22] The biggest problem of those early attempts is that they all require a large degree of human intervention and none of them were capable of increasing their performance from learning.

The milestone that heralded the transformation from semi-automated to almost full automated recognition is the development of a system called Eigenfaces in 1988 by mathematician Michael Kirby and Lawrence Sirovich, who applied principal component analysis (PCA), a linear algebra technique, to represent faces by a "relatively low-dimensional vector" and identify human faces through their deviations from the average.[23] Three years later, the Eigenface approach was further expanded by Matthew Turk and Alex Pentland at MIT, opening the door to the first instances of automatic FRT.[24] Since then, the interests in developing and utilizing FRT had grown steadily and significantly, with a variety of novel algorithms, large competitions, and commercially available systems being crafted. Some important fruits during 1990s include Elastic Graph Matching (EGM) that recognizes faces based on a graph representation of faces extracted from other images, Local Feature Analysis (LFA) that gave rise to the famous FaceIt system of Visionics company, and the Face Recognition Technology (FERET) program initiated by the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST), who created one of the first face database and held several large competitions, leading to a boom in the market at that time. In the 2001 International Conference on Computer Vision (ICCV), Paul Viola and Michael Jones from Compaq Cambridge Research Laboratory showed a quasi-real-time face detection system, which could detect at 15 frames per second.[25]

In recent years, with the revival of AI and the development of novel machine learning techniques such as support vector machine (SVM), artificial neural networks (ANN), convolutional neural networks (CNN), and deep learning (DL), FRT gains fresh momentum, and a growing attention was directed to FRT in unconstrained scenarios. Databases like Labelled Faces in the Wild (LFW) were created to facilitate FRT research and tests. Today, the mainstream technique of FRT is using DL and big data, with ever deeper ANN and ever larger data volume. For instance, DeepFace, the DL FRT of Facebook was trained on 4 million images uploaded by Facebook users, with a near-human accuracy of 97%, while FaceNet, Google's 99.63% accurate CNN FRT, was trained on 200 million face images of 8 million people and has 22 layers.[26] Machine learning algorithms allow computers to learn from a large amount of training data, using techniques like gradient descent and heuristic search to gradually tune the weight of each node in ANNs until the optimum results are reached. It provides a way for the algorithms to learn from experience with minimum human intervention.

Nowadays, FRT is utilized in a wide range of areas. Compared with other biometrics such as fingerprint, FRT does not need physical contact to function. FRT can also operate from far away with little awareness, so it naturally fits in surveillance and security tasks. Some systems are surprisingly efficient. For example, China has deployed a powerful CCTV network across the whole country, which caught the BBC reporter John Sudworth in seven minutes.[27] FRT could also be used to verify people's identity in many situations, such as banking, online shopping, unlocking cell phones, and passports. For example, in September 2017, Alibaba in Hangzhou, China, launched the "smile to pay" service that allowed people to make a payment in KFC by canning the faces. Some public

bathrooms in Beijing even use FRT to verify people's identity in order to prevent toilet paper theft.[28] FRT could also be used in Human-Computer Interfaces (HCI) for many purposes. For example, Coca-Cola used FRT on interactive vending machines across Australia to detect customer interaction.[29]

2. What's in a face?

The practice of classifying people with tags has a long history. Since the dawn of human civilization, people have been organized into different hierarchical categories based on their physical characteristics such as gender, age, skin, and color, which are associated with stereotypical images and discrimination. Were Solomon Northup not an African-American, the odds of him, as a free citizen, being kidnapped and sold as a slave would have been much lower.

Among all the physical traits, the most visible and natural one is the face. A face is not only an indication of our identity, but also a portal to everything about us, such as emotions, ethnicity, personality, affection, and other personal information. Reading faces makes up a significant portion of human's social life. The exploration and practice of judging people from their facial features, or so-called physiognomy, can be traced back to ancient China and Greece, where it was widely believed that certain traits of a face could tell a person's character, destiny, health, wealth, background, and other things. In *I Ching* (*Yijing*), also known as *Book of Changes*, the oldest Chinese divination text, physiognomy is one of the Five Techniques. In western culture, Aristotle believed "soft hair indicates cowardice," and a broad nose was "a sign of laziness," like in cattle. And it is said that Pythagoras even selected students based on their facial features. Another example is Liu

Bei, the founder of the state of Shu Han in the Three Kingdoms period of China, who was believed to have many facial features that imply great achievements in the future such as big earlobes that can be seen by himself, which even helped him win trust from ordinary people long before his political career.

Today, although physiognomy has long been refuted as having a concrete scientific foundation, links between facial features and personal characters are still widely believed to exist. It is evident that facial appearance and personality can influence each other, and they can also be affected simultaneously by external factors like environment as well as internal factors like hormonal levels and genetic expressions. For example, the prenatal hormone theory (PHT) is a widely accepted theory that associates facial appearances with sexual orientation with fetal androgen signaling on sexual differentiation. Dozens of facial metrics were found to be of significant difference between homosexual and heterosexual people, for instance, the puckers on mouths, the size of foreheads, the turning-up of noses, and the depth of chin. Carmen Lefevre, Gary Lewis, and others found that higher levels of testosterone are positively associated with wider faces and bigger cheekbones, and these people—no matter male or female—tend to have male behavioral traits, such as "aggression and status-striving."[30] In addition, Benedict Jones at the University of Glasgow thinks people with thinner faces have a smaller chance of infection, arguably because the accumulation of fat in the upper part as oppose to the lower part of the body may indicate less health. Also, red facial coloration is evidently believed to be a sign of good circulation and women's fertility.[31] There are even studies on whether people can accurately assess the intelligence of others merely by watching their faces. Karel Kleisner of Charles University in Prague and colleagues concluded from an experiment that the IQs

of men could be accurately perceived merely by showing their face photos to other people, while interestingly, no significant correlation found between women's perceived IQs and their faces.[32] Similarly, many studies have been conducted, attempting to find links between facial features and underlying information like political stand, self-esteem, sexual orientation, the quality of relationship with their partners, and so on, all of which assume psychological states are rooted in and have an impact on biological characteristics.

Despite correlations found, however, there are some problems with these links. With few exceptions, most facial features in these studies were perceived by humans, but human perception is not accurate and explainable. For example, a study that shows by merely looking at women's "emotionally neutral" faces, men can tell who is wearing unseen but self-reportedly more attractive clothes.[33] This could be easily explained by the unconscious relation between facial expression and psychological states like confidence, but which features are important and how other people perceive them are largely unknown.

Therefore, with the huge potential of pattern recognition techniques, some researchers started to use computers, specifically, AI techniques in recent years, to look for subtle facial features that may imply characters. For example, the researchers in the Department of Psychology at Brock University, Canada, use logistic regression to pinpoint facial metrics that are different between homosexual and heterosexual people, and they also use PCA and discriminant function analysis (DFA) to examine linear combinations of facial metrics.[34] The Stanford study mentioned in Chapter 1 used Deep Neural Network (DNN) to extract facial features and logistic regression to make further predictions. Similarly, the criminal-prediction study conducted by Shanghai Jiaotong University used machine learning techniques like logistic regression, SVM, CNN, and k-Nearest Neighbors

algorithm (KNN).

3. How does Facial Recognition Technology recognize and classify people?

The two types of FRT application, identification and classification, both contain two common steps—face detection and facial feature extraction. The next steps for identification applications are face-matching and providing the recognition result, while the next steps for classification applications are face-classification and outputting classification results.

For the first step, detecting faces in images, two approaches dominate: feature-based geometric approach and image-based photometric approach. The former had been the focus of the field until the mid-1990s. In geometric approach, facial features are extracted based on the positions, sizes, shapes, and other traits of facial components like the eyes, mouth, and ears.[35] Components can be extracted by detecting their edges, out of which the feature vectors can be built. Sometimes, feature blocks are distinguished by dividing face images into different regions based on grayscale difference before any specific features can be identified. Since 1997, machine learning techniques started to enter the field and show significant promise, allowing algorithms to be trained on large numbers of examples to learn how to detect facial components with high accuracy, for example, Sung and Poggio's mixture of Gaussian model, Osuna's support vector machine approach, Rowley's neural network approach, and Roth's Winnow learning procedure. In recent years, methods for detecting face images in unconstrained scenarios are in active research, such as detecting faces with a rotated angle and 3D views.

After a face is detected, the face image is normalized, and features of the face will

be extracted and serve as input data fed into face databases or classification systems for further analysis. Depending on the kind of classification task, the features that need to be extracted can be different. For instance, they could be features like eyes, mouth, and ears, or local features like lines or fiducial points.[36] Sometimes the process is conducted at the same time as face detection. There are three main methods for facial feature extraction: generic methods that focus on traits like edges and curves, feature-template-based methods that detect features like eyes and mouth, and structural matching methods that consider geometric constraints of the features. Take the first method as an example. Hua Gu, Guangda Su, and Cheng Du from Tsinghua University described a corner detection approach to locate the important feature points of a face. First, eyes are located by searching for valley points of luminance in the upper area of face images and calculating the degree of symmetry of two eyeballs. Then, the location of the nose is found by locating the vertical highlight area between the two eyes, after which, the nostrils are located by searching for valley points on both sides of the lower area of the nose. After that, the algorithm searches for a dark area below the nose to locate the mouth, while eliminating the beard. Then, the values of the features are stored as "faceprint."[37] The template-based approach, just as its name implies, analyzes face image based on a template with predefined parameters. Combining with statistical models and machine learning techniques, the performance of facial feature extraction is enhanced significantly. For example, T.F. Cootes, G.J. Edwards, and C.J. Taylor proposed the Active Appearance Model (AAM), which trained on 400 face images, each of which was labeled with 68 landmark points generated by applying PCA.[38] A recently widely used facial-extraction is VGG-Face, a DNN trained on 2.6 million images.[39]

In recognition applications, after the facial features have been extracted, the system searches the features in the database for a match. Basically, the system compares the features with existing "faceprint" in the database with two possible procedures—identification and verification. Identification is to find out who the person is, while verification is to pull out the faceprint of the target person in the database to confirm whether the captured person is who he/she claims or is claimed to be. The verification procedure outputs a matching score through a function that measures the similarity between the feature vectors. If the score exceeds a predefined threshold, the claim can be seen as true; otherwise, false. The identification procedure outputs a set of matching scores through the same function. Among the ones whose scores exceed the threshold, the one with the highest score is most likely to be the identity of the person captured in camera.[40]

In classification problems, the techniques used to label people belong to the problem set called classification or clustering, based on whether supervised learning or unsupervised learning is in question. The difference between supervised and unsupervised learning is the training data of the former is manually labeled by humans. Classification is to let algorithms learn to categorize training data in order to decide into which category new data falls. A classification algorithm is usually called a classifier. Several examples of classifiers are the linear classifier (logistic regression), neural network, support vector machine, random forest, and nearest neighbors. For a neural network classifier, there are layers of largely homogeneous nodes, or artificial neurons. The layers between the input layer and the output layer are called hidden layers, which define the depth of the network. Each node is connected to every node in the neighboring layers with a weight. The weighted input sum of a node, together with the activation value calculated through the

activation function, determines whether the node “fires” or not. The training data is fed into the input layer, traveling through the hidden layers, and finally outputting a category. The output data is then compared with the labeled category. If the result conforms with the label, say, “homosexual,” the weights of the nodes are strengthened; otherwise, the weights are tuned again and again, until the optimum result is reached. In this way, a classifier that could accurately predict the labels of the training data is developed and ready for new data. For example, a gender classifier will be fed with face images labeled with “male” and “female.” While at training, if a male picture gets a “female” output, the weights will be tuned through techniques like gradient descent, until the output conforms to the label.

CHAPTER 3

ETHICAL ISSUES OF CLASSIFICATION FACIAL RECOGNITION TECHNOLOGY

1. Ethical frameworks

Ethics, or moral philosophy, is the norm that helps people distinguish right from wrong. In other words, it deals with what ought to be done, instead of what it is, serving as principles or guides for individual and collective behaviors. There are other normative subjects that guide human behaviors, such as law, religion, and etiquette. Although these subjects share many common properties and their origin and historical development overlap to some extent, ethics has many unique traits and is radically different from the rest of them. For one thing, its coverage and punishment differ from laws. It also requires more rational reasoning instead of unfalsifiable authority to enforce and justify than religions do.

In western culture, the term ethics derives from *ethos* in Greek word, meaning "habit, custom." In Chinese, although "*伦理*"(ethics) and "*道德*" (morality) are used largely as synonyms, it is interesting that the two actually have subtly different meanings and origins. "*伦理*" literally means "the laws of relations," focusing on the should-be relationships between individual agents. "*伦*" originally means natural orders and the hierarchical relationships derived from the natural orders. It came from the "Five Relationships" described in Confucianism—the monarch-subject relationship, the father-son relationship, the husband-wife relationship, the relationship between brothers, and the relationship between friends.[41] On the other hand, "*道德*" (morality) deals with the values and norms of conduct from the perspective of the whole society. The term derives from the

Tao Te Ching (*道德经*) by Laozi, in which, "*道*" means the true nature of the universe, and "*德*" means the way of doing things according to the law of "*道*."[42]

In general, the theories concerning ethical issues can be organized into a hierarchical structure based on their objects of study. At the highest level is the area called meta-ethics, which explores the basic meaning and nature of ethical statements and how they are supported or defended. In other words, it is the ethics of ethics. There is a collection of areas called applied ethics at the bottom of the hierarchy, concerning with particular "moral problems, practices, and policies in personal life, professions, technology, and government."[43] In between is the normative ethics that seeks to establish a self-contained theoretical framework for analyzing ethical issues. This paper focuses on the latter two, trying to pinpoint the particular ethical issues related to the FRT-driven classification tasks and explore how to incorporate normative ethical frameworks into the analysis.

Three main types of frameworks can be identified in normative ethical theories—utilitarianism, deontological, and virtue theory, focusing on actions, consequences, and characters respectively.[44]

Utilitarianism looks into the consequences of actions, seeking to produce "the greatest good for the greatest number."[45] The "greatest number" is easy to quantify, but the "greatest good" is difficult to define. That's where debates surround. The hedonistic utilitarianists, such as Jeremy Bentham, believe that physical pleasure is the only good thing that is worth pursuing, while pain the only evil that must be avoided. On the other hand, eudemonistic utilitarianism distinguishes happiness that requires "higher faculties," as John Mill put it, from mere sensorial pleasure, and regards the former as more valuable even if it means suffering from acute pain or it might seem to be smaller in quantity.[46] In

this sense, the people of Zion that fight against the Matrix are better and happier than the ones living in the simulated reality even though the latter has a much higher degree of sensorial satisfaction. Another type of utilitarianism, which focuses on rules rather than individual actions, has supporters like John Hospers. According to this rule-utilitarianism, the rightness of an action is judged by the "consequences of its universalization" rather than the immediate consequence of the action itself. For example, deploying an FRT system on M street probably will help capture some thieves, but "track every person on M street without consent" might be a bad rule to follow if universalized, which could lead to adverse privacy consequences.

While utilitarianism concentrates on the outcomes of particular actions or rules, the deontological ethical frameworks look into the intrinsic features of behaviors to estimate whether they are inherently right or wrong. The greatest deontologist in history is Immanuel Kant, who proposed the important concept of categorical imperative, or "the ultimate criterion for the moral acceptability of all actions."[47] He believed that an action is good only if it conforms with a maxim that should be unquestionably universalized. In addition, he opposed treating people, including yourself, merely as a means, so he considered lying unacceptable since it ignores people's goal and values, and deprives people's right to make autonomous decisions. In other words, Kant emphasized the importance of respecting people as autonomous agents, which later becomes a common ethical principle in a variety of areas, such as the Belmont Report, which was developed in 1979 to protect human subjects in biomedical and behavioral research.[48] Also, the four principles of computer ethics developed by Norbert Wiener—the principle of freedom, the principle of equality, the principle of benevolence, and the principle of minimum

infringement of freedom, could also be seen as an incarnation of Kantian formula of humanity. In Kant's spirit, treating human faces as a mere mean to predict criminal behaviors or classify people into different groups only for the convenience of administration without considering their goals, values, emotions, and interests is morally wrong. Also, according to deontological point of view, ethnical discrimination in airport security systems, for example, the requirements for Muslims to have more strict screening, is morally wrong, even though some people believe it would prevent terrorist attacks (teleological approach).

The third ethical framework is virtue theory, the oldest ethical theory in the western culture that could be traced back to Plato and Aristotle in the 5th and 4th centuries BCE. This framework focuses not on behaviors, rules, duties, or consequences, but on the characters of actors such as honesty, patience, courage, and other traits that could promote excellence. Excellence is the very meaning of the etymology of "virtue" and is deemed to be able to ultimately lead to human flourishing, for right actions are believed to be generated effortlessly and inevitably from good characters. In other words, authentically good behaviors stem from virtuous characters, not from fear of religious punishment, legal sanctions, peer pressure, or other things described in Michel Foucault's reflection on Jeremy Bentham's notorious design of the Panopticon.[49] FRT systems, however, especially when incorporated into surveillance systems, have the potential to work against the cultivation of virtue. "Surveillance technologies that work too well in making us act 'rightly' in the short term may shortchange our moral and cultural growth in the long term," as Shannon Vallor put it in her book *Technology and the Virtue*, because the seemingly good behaviors come from the belief that some Big Brother is watching, not from the

conviction of the rightness of the actions. One can imagine, if a total surveillance society with FRT is formed, once the system is deemed overthrown, the social order could barely be maintained, let alone the motivation of individuals to behave morally.

These frameworks have been woven into numerous social practices, among which, the institutional review board (IRB) is one of the most prominent. IRB is a type of committee that oversees whether research that involves human subjects, such as biomedical studies, is "carried out in an ethical manner"[50] and whether the involved human subjects are properly protected from various kinds of harms. IRB originates partially from the *Belmont Report* issued in 1978, whose full title is *Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Three basic ethical principles were proposed in this report: (a) Respect for Persons, which requires the acknowledgement of individual autonomy and the protection for those with diminished autonomy; (b) Beneficence, which ensures subjects' well-being by minimizing possible harms and maximizing possible benefits, and (c) Justice, which emphasizes the fairness of distribution of both the benefits and burdens of the research.[50] The first principle perfectly reflects Kant's formula of humanity, while the second conforms with the duty of non-maleficence, one of the moral requirements proposed by William David Ross (1877-1971), another deontologist.[51] The justice principle accords with the utilitarian principle of producing the "greatest good for the greatest number." In 2012, *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research* was released by the U.S. Department of Homeland Security, in the hope to apply the three principles of *Belmont Report* in the emerging fields involving information and communication technology (ICT). A fourth principle—Respect for Law and Public Interest, was added to the list, in order to keep pace

with the development of technologies that might involve issues like privacy, confidentiality, and informational integrity. Despite these efforts, however, the guidelines sometimes are criticized for being vague, outdated, and not able to cover all the emerging technologies. For example, the Stanford gay detector study was conducted under IRB review, thus, data ethicist Jacob Metcalf criticized that IRB is lack of "consistent standards or transparent review practices," and the guidelines are "outdated and often irrelevant" because many rules were developed forty years ago with the attempt to monitor very specific type of research that is inapplicable for data science.[52]

2. The ethical issues of discrimination

While there are a lot of ethical issues associated with FRT, this thesis mainly focuses on issues of discrimination that may be caused by FRT while doing social sorting tasks.

What is discrimination? According to the Cambridge Dictionary, discrimination is treating certain people "in a way that is worse than the way people are usually treated."[53] Ambiguity exists, however, in this definition. For example, what is the way people are usually treated? What does "worse" mean? Who are those certain people? There are also many types of discrimination that seem unproblematic, such as price discrimination in economics, with which Georgetown students could get special discounts in certain apparel stores on M. Street. Are people without Georgetown ID treated in a way that is worse than the way Georgetown students are treated? Arguably yes. But does this discrimination raise ethical issues? Probably not. Therefore, discrimination needs to be carefully defined in this context.

Theoretically, discrimination is treating people differently based on the differences

of some of their traits. Some types of discrimination raise no obvious ethical issues, while others do. The latter types are the focus of this thesis. One of the fundamental differences between the two is that the latter type of discrimination is based on the "membership of a socially salient group,"[54] such as gender, age, race, religion, and sexual orientation, as reflected as "protected groups" in U.S. federal law. For example, if a male applicant is favored over a female applicant by an employer because the former applicant has longer experience in the industry, it may not be seen as discriminatory. But if the two applicants have similar background but somehow the male is favored anyway, it is highly possible that the recruitment decision is made simply because of their genders. Then it may involve discriminatory action, because "women" are a socially salient group, let alone they are a historically discriminated and oppressed group in job market.

Therefore, discrimination can be defined as the disproportionate adverse impact due to differentiated treatment to a socially salient group. Many people believe discrimination is morally wrong. But what's wrong with it? I'll discuss the wrongness of discrimination from the three frameworks mentioned in the previous section—utilitarianism, deontology, and virtue theory.

From utilitarian point of view, a morally righteous behavior should promote "the greatest good for the greatest number." Discrimination may undermine the ability of either individual or the society as a whole to create the greatest good for the greatest number, because it is harmful to the person involved, lowers welfare for society as a whole, and worsens the situation of everyone off. For instance, consider a situation where a local bank uses a piece of software that includes an FRT component to help make loan decisions for small businesses in an area where, historically, women were less engaged in business than

in domestic work. By learning the files of past loan applicants, the software comes up with a black-boxed profile for what kinds of applicants are most likely to pay off the loan. Since there are few females in the training data, the FRT component might conclude that people with masculine facial features are more likely to pay off the loan, therefore rejecting most women applicants. This software definitely discriminates against women and has harmful consequences. Individually, it relentlessly reduces each woman's opportunity to get funded for her business, jeopardizing her chance to professional success and the financial prospect of her family—think about the possibility that she is a single mom who works hard to raise her kids. Collectively, this software deprives the economic, social, psychological, and many other advantages that could have been benefiting the whole society if a reasonable number of female entrepreneurs get grants. Therefore, discrimination is harmful in terms of utilitarianism, and FRT with social sorting purposes surely can be used in a discriminatory way that hurts both individual and the society.

Deontologists emphasize the significance to treat people as autonomous agents, and opposition to treating people as "mere means." Everyone should be treated with equal respect, for everyone has her own goals and values that are worth respecting. Discrimination by definition treats people unfairly; if on the basis of "irrelevant characteristics," as Lena Halldenius put it,[54] it will awfully undermine social equality. If the decision to treat people differently is based on the grounds that are irrelevant in the context, then an unfair, wrongful discrimination can be identified. For example, lower limb disability is irrelevant in the context of academia but could be relevant in the context of pilot recruitment. Therefore, a rejection decision on the basis of this disability made by a graduate school is discriminatory while the same decision made by an airline company is

not. Some people may argue that disability is relevant in graduate school because a might need a special facility to get into the classroom. As long as the facility in need is not astronomically expensive, however, and her academic success is largely irrelevant to the disability—in other words, she has no serious psychiatric disorders that may endanger the safety of her fellow students—the graduate school should not reject her application based on the disability. Discrimination based on irrelevant factors, however, exists widely in the society. For example, a correspondence test conducted in the Chicago labor market found that resumes with Anglo-Saxon names get one third more call-backs than identical resumes with non-Anglo-Saxon names.[55] This is obviously discriminatory because names, or the ethnic groups indicated by names, are irrelevant to job performance. The irrelevant criterion is particularly important to the analysis of FRTs because facial features are largely irrelevant in many social contexts. Treating people differently on the basis of their perceivable or unperceivable facial features could lead to acute discrimination. For example, sexual orientation might be relevant to romantic relationships but is largely irrelevant to job performance, academic achievement, dietary habit, criminal tendency, and many other things. If someone is treated differently in a way that worsens her situation, for example, getting fired by a company or rejected for research funding, simply because she is a lesbian, it is seriously discriminatory; for sexual orientation is relevant to neither her job performance nor her research skill. So, an FRT that classifies people according to their sexual orientations surely has the risk of contributing to immoral discrimination. Such discrimination could result in unfair distributions of wealth, income, or positions.

Virtue theory focuses on cultivating the virtuous characters such as honesty, patience, courage that promotes excellence. Discrimination hurts the virtues on the levels

of both individual and community. On the individual level, as J. L. A. Garcia analyzes, both the discriminators and those who suffer from discrimination face challenges to live virtuous lives. The discriminator may be burdened with moral condemnations and reluctance to accept the fact that they are no different from the people they held in contempt, for example, the "burden of white privilege." Meanwhile, those who face discrimination may suffer from psychological issues such as low self-esteem and antisocial personality, which hinder them from developing truly autonomous, virtuous characteristics that lead to good behaviors.[54] For the community, discrimination may lead to greater homophily and ultimately segregation in social networks, diminishing democracy and freedom. Countless lessons can be found in the history of U.S, South Africa, and many other countries. The virtue of the whole society may be jeopardized.

3. The discrimination issues of classification Facial Recognition Technology

As mentioned in the last section, discrimination can be defined as the disproportionate adverse impact due to differentiated treatment to a socially salient group, and it could lead to serious ethical issues. But why do FRTs with social sorting function have a high risk of resulting in discrimination?

First of all, FRT is inherently biased due to the training data. Although machine learning algorithms are able to infer to some extent, they fail to incorporate human-level common sense in their reasoning. This is partly because the mechanism of common sense is not well-understood yet. Today's machine learning algorithms, however colossal or smart, are only as good as the data on which they are trained. So, the biases within data are

easily hard-wired into the models through data-mining. In other words, if the training data is biased in some way, the FRT will give results that are biased in the same way. In May 2014, the White House report *Big Data: Seizing Opportunities, Preserving Values* was among the first reports that mentioned the potential discrimination in big data. It is reported that "big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace."[56] For example, a research conducted by Joy Buolamwini, a researcher at MIT Media Lab who is mentioned in Chapter 1, found that FRT is 35% less accurate when working on faces of females of color than white males.[57] There surely are some technical factors, but one of the core reasons is that the algorithm is trained mainly on white males. For example, it is estimated that in one widely-used FRT training dataset, 75% are male, and over 80% are white. Even if there are no pre-assigned parameters related to race and gender, the biases tend to be hard-wired into the algorithm through the training process, working as a source of discrimination. For instance, an FRT used for security-check at an airport might spend more time struggling with checking the identities of black women than white men, treating black women "in a way that is worse than the way people are usually treated" on the basis of their features as a socially salient group—women with dark skin. This treatment comes from a factor that is irrelevant in the context—skin color and gender have nothing to do with an individual's chances of carrying illegal items on the flight. Also, this wasting of time and resources of both black women and the security staff worsens everyone off and increases inequality of opportunity. And legal scholars like Solon Barocas and Andrew D. Selbst argue that the laws right now "largely fail to address" this discrimination from data mining.[58]

Second, FRTs with social sorting functions could be easily used to develop applications that classify people, intentionally or unintentionally automating the behaviors of discrimination. Imagine how the Stanford gay recognition algorithm could be used for homophobic, discriminatory purposes in social contexts that are irrelevant to sexual orientations. Similarly, if a criminal prediction FRT component is integrated into an auto-gate of a metro station, which prevents people labeled with "high chance to break the law" to enter the metro even if they've never done anything wrong, it could raise huge concerns of 1984-ish totalitarian scenario.

Third, FRTs, especially those with social sorting functions, may increase the number of discriminatees. Three kinds of discriminatees can be identified according to their historic status—classical, standard, and novel forms. Classical discriminatees are the groups that fought for their civil right in the 1960s, such as African Americans; standard discriminatees are those groups that are not in the center of discrimination discourse initially, such as sexuality, age, and disability. These two made up most of the protected groups in U.S. federal law. The Novel forms of discriminatees are groups that have not been "commonly recognized," such as obesity, lookism, and transgender.[54] FRT could increase the number of discriminatees for the former two types of groups, and add new groups into the third type. On the one hand, many socially salient groups, especially the "protected groups," could be easily identified through facial features, such as gender and race. In many social contexts, however, even though those features are noticeable to naked eyes, the memberships of those groups are often ignored because of the irrelevance. The potential automation of the classifying tasks of FRTs can automatically add those dimensions into decision-making processes, potentially discriminating against those who

would not have been discriminated otherwise. On the other hand, FRT may add new groups to the list of groups of discriminatees, creating new forms of discrimination. For example, sexual orientation usually can't be easily identified through naked eyes, but with FRT such as the Stanford gay recognition algorithm, it can be predicted with arguably high accuracy. Such algorithm could enable business or government agencies to refuse to provide services to homosexuals, creating new forms of systematic discrimination against the LGBTQ community. A recent hot debate is whether the Tastries Bakery owner Cathy Miller has the right to refuse to make a cake for Eileen and Mireya Rodriguez-Del Rio's same-sex marriage wedding because of her Christian faith against homosexual behaviors. Although Kern County Superior Court Judge David R. Lampe ruled for Miller, it was only based on his belief that creating a cake is "an act of artistic expression," which is protected by the First Amendment.[59] Most commercial services are not artistic expressions, for example, restaurants, shopping malls, grocery stores, and car dealers. If gay-detecting FTR is deployed in such business settings, they could distinguish people based on their sexual orientations and treat them differently, for example, charging people of certain sexual orientations with a higher price, or providing differentiated products or services, which could be discriminatory.

Furthermore, the potential discriminations of FRT are difficult to discover and mitigate because the algorithms are mainly black-boxed. On the one hand, deep learning, the mainstream technique for FRT, is a connectionist model. That is to say, there is minimum semantic information built into the algorithm. So, how the hidden layers are connected is largely black-boxed, making it difficult to articulate how and why the network makes such predictions, not to mention the difficulty of pinpointing problems and dealing

with them. The results depend on the labels of the training data. On the other hand, machine learning draws associations among factors in the sense of statistics instead of the relevance of the context, which is more nuanced and subtle. Therefore, it ignores the "irrelevance" criterion mentioned in previous sections. For example, lip curvature and eye inner corner distance are irrelevant to a person's behavior at a shopping mall. The criminal prediction model developed by researchers at Shanghai Jiaotong University, however, drew a connection between these two factors and the person's possibility of being a criminal. If a shopping mall uses this algorithm to decide who is allowed to shop in the mall, then it could refuse some customers because it infers that a person labeled with "high chance to break the law" will probably conduct disorderly behaviors in their property. The association is hard to discover by the external world. Thus the discrimination caused by it is hard to reveal. Also, since the biases are the result of past data, it is usually difficult to find a method to adjust the historical data.[58]

CHAPTER 4

DESIGNING LESS DISCRIMINATORY FACIAL RECOGNITION TECHNOLOGY

In this chapter, I'll give some suggestions to address ethical challenges in the design and implementations of both FRT practices and FRT research. It is important to make the distinction between practice and research, because they differ in many ways, such as purpose, methodology, and assessment, even if they involve similar ethical principles. According to the Belmont Report, in biomedical research, practice means the diagnosis or treatment specifically provided to individual patients in order to increase his/her well-being but no one else, while research is designed to test a hypothesis in order to "contribute to generalizable knowledge." Similarly, FRT practice is to provide commercially available products in a pre-assigned context. For example, Megvii Face++, a Chinese FRT company, develops an FRT system for e-commerce company Alibaba that allows customers to make payment by simply smiling in front of their cell phone camera.[60] By contrast, FRT research is to develop advanced algorithms, or to use FRT to test a hypothesis in any area. For example, one of the purposes of Stanford gay recognition research is to use FRT to test the prenatal hormone theory of sexual orientation.

However, it is worth noting that the distinction between the two is gradually blurring. For one thing, product development always comes with research. More and more research is carried out by commercial companies because research plays an important role in terms of financial return. Also, in the age of information technology, the private sector has access to large amounts of data, which is of high academic value, and which could provide insight into understanding human society. So, research is encouraged in many

companies, such as Google, Facebook, and Microsoft. Many tech giants have built their own research institutes. Meanwhile, many academic studies are conducted in the consideration of or contribute to commercial or governing practices. This blurring of boundaries complicates the situation because companies usually don't rely on an IRB procedure, which governs the ethics of academic research, to review their R&D. This lack of an IRB could lead to serious issues. Therefore, it is important to explore some universal principles and procedures that are applicable either for practice or research.

For FRT practice, I propose using a value sensitive design approach to investigate and mitigate the discrimination issues, while for FRT research, I recommend extending IRB and the subsequent Menlo Report, in order to accommodate new issues in FRT. In the new context where practice and research are hard to distinguish, I use Facebook as a case study to demonstrate how to evaluate the potential discriminatory risk of FRT.

1. Value sensitive design in Facial Recognition Technology industrial practice

What is value sensitive design?

Value sensitive design (VSD) is a design approach that aims to incorporate human values into the design process of a particular technology "in a principled and comprehensive manner."[61] In other words, VSD intends to build a consistent value system into technological applications before they could be deployed in the hope of mediating the interests of different stakeholders, avoiding value conflicts, and promoting moral values. This approach has earned much attention among designers, especially when designing information systems.

Human value refers to the things that people consider important, such as financial benefit, privacy, and autonomy. To determine what values matter, four principles from the writings of Norbert Wiener, one of the pioneers of Computer Ethics can be used as a guideline, including the principle of freedom, the principle of equality, the principle of benevolence, and the principle of minimum infringement of freedom.[62] This paper follows the principle of equality, in particular, the freedom from discrimination, including not to be discriminated by others, not to discriminate others, and not to help create a discriminatory discourse or environment.

The methodology of VSD consists of three parts—conceptual, empirical, and technical investigations (Friedman et al., 2008). Conceptual investigations involve the identification of direct and indirect stakeholders of the technology, how they are affected by the design, what they value, how different values compete, and how to weight them. Empirical investigations seek to answer better key questions raised in conceptual investigations using quantitative and qualitative research methods that are common in social science, such as interviews, survey, content analysis, and experiments. For example, to explore how stakeholders prioritize competing values of social media—say, privacy and convenience—interviews can be used to collect their opinions. Control groups and experiments could also be used to see whether there is any improvement in value building after people use such designs. Finally, technical investigations focus on how various technological factors influence the values identified in conceptual investigations, and how to design those factors in a way that supports the values in question. [61]

For example, in designing firefighting equipment, if gender equality is identified as a key value in female firefighters, then the weight of hoses could be recognized as a

technical factor that influences the accessibility of the equipment to general female users. Using interview, the empirical investigation may show that female firefighters feel frustrated if the hose is too heavy for them to lift and carry, affecting their job performance so that many of them chose to leave the department, which could undermine gender equality. So, a designer who appreciates this value might consider designing hoses in a way that is easy for a general woman to carry, say, decreasing the diameter. However, a competing value may raise—the efficiency of firefighting. The decrease of the hose diameter reduces the rate of flow, which could jeopardize the efficiency of fire extinguishing. Further investigations could be conducted to see how different stakeholders prioritize the values of gender equality and work efficiency, and then, further technical iterations can be carried out, hopefully being able to design a hose in a way that aligns with the ideal value system, say, by using more light-weighted material instead of decreasing the diameter to maintain the flow rate while reducing the weight at the same time.

Thus, VSD could help designers and engineers build systems that are more beneficial not only to the users of the technology but also to the value system of the whole community and society.

Value sensitive design in FRT to avoid discrimination

As discussed in Chapter 3, FRTs with social sorting functions have a high risk of causing discriminatory issues. So, freedom from discrimination should be considered as an important value when designing FRT systems. Here, I propose some detailed suggestions that can be integrated into standard VSD procedure when designing FRT systems.

Conceptual Investigation. First, designers should identify the usage scenario of the FRT and find out the direct and indirect stakeholders in it. For example, for an FRT

used in airport security system, the direct stakeholders could be the security staff and passengers, because they interact with the camera and monitor directly, while the indirect stakeholders could include other staff and the management department of the airport. In particular, since we are focusing on discrimination, potential discriminators and discriminatees must be identified as subgroups of both direct and indirect stakeholders as well. One tricky aspect of machine learning system in general, and FRT in particular, however, is that the inherent biases are difficult to identify. For this reason, some tools for algorithmic auditing are under developing.[63] For example, Sarah Tan and colleagues developed a "two-pronged approach" to find biases in machine learning algorithms. The first part simulates the algorithm being examined and gives a "risk score" on the basis of one sub data set from the initial data, while the second part is trained on real-world data, in order to determine the parameters in the initial data that matters most to the outcomes.[64] Using this model, They found that the COMPAS algorithm of Northpointe company, which is used by courts to predict the recidivism risk for defendants, may bias against some age groups and racial groups, despite the company's claim that it is race-blind.[65] This result agrees with an earlier investigation by ProPublica.[66] Using this kind of tool, designers can identify the potential discriminators and discriminatees as subgroups of stakeholders. For example, as Buolamwini found out, FRTs are 35% less accurate on black women than white men. So, if one such FRT is to be used in airport security, black female passengers must be identified as an important group of direct stakeholders, and their opinions and values must not be ignored during the development of the FRT system. Not all biases, however, would necessarily lead to wrongful discrimination, as discussed in the previous chapter. A careful investigation must be conducted to see whether the biases are relevant in the context. For

example, if gender biases are detected in an FRT, it could lead to discrimination in a bank system that approves credit card applications, but it doesn't necessarily result in discrimination in an online product recommendation system, since gender is irrelevant in the context of credit card, but could be highly relevant in the context of online shopping.

Then, other values must be identified, such as privacy, efficiency, data security, and psychological well-beings, in order to look for complementary and conflicting relationships among values. Other conceptual questions include how different stakeholders engage with the values in question, how they might weight different values, what the benefits and harms for each group of stakeholders are, how value conflicts between groups as well as within groups, and so on. For example, for an FRT that recognizes gays in shopping malls, designer should ask how homosexual people might engage with the system, how the system might influence the sellers' decision-making towards homosexual people, and how these decisions might affect the value of freedom from discrimination, equality of opportunity, and privacy of the clients.

It's worth noting that the relationships between values may not hold universally to different stakeholders. Two values that are competing to one stakeholder may compensate each other to another. For example, a police officer on patrol equipped with a body camera with gay-recognition FRT algorithms may believe that knowing a person's sexual orientation before any action is carried out could help them make more reasonable decisions when he pulls over someone's car, but it may lead to differentiated treatment of homosexual people. In other words, to a police officer, the value of freedom from discrimination conflict with the value of autonomy in some sense. To the person whose face is analyzed by the body camera, however, concealing her sexual orientation is her

freedom, and could help her stay away from certain discrimination, so, the value of freedom from discrimination compensate the value of autonomy.

As we can see, the value system and stakeholders often form a complicated network, so, here I propose using some network visualization technique to map out the relationships between people and value. See the example in Figure 1, which is a network graph that

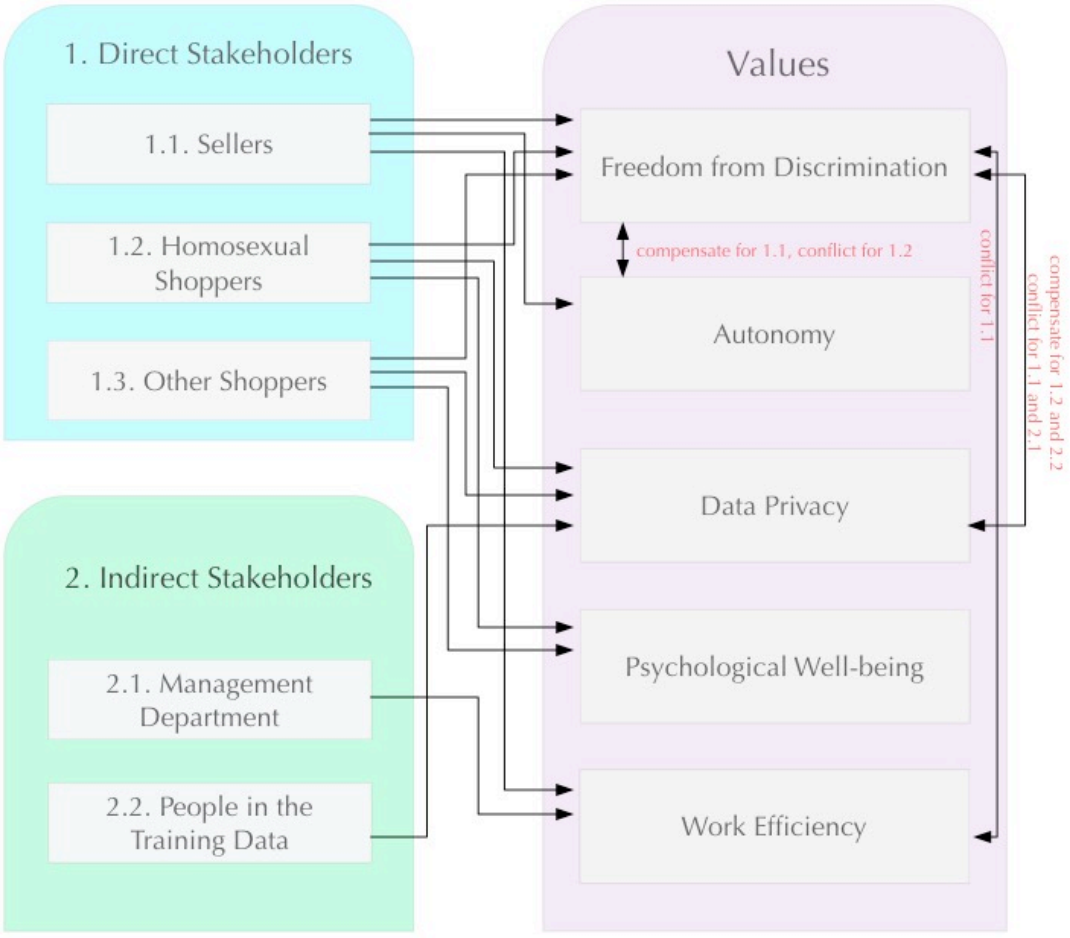

Figure 1 An example of the network of stakeholders and values in a product recommendation system with gay-recognition FRT in shopping malls.

shows the relationships between stakeholders and values of a product recommendation system in shopping malls with gay-recognition FRT algorithms. This graph is inspired by

the affiliation network used in social network analysis (SNA) to show the affiliation relationship between people and organizations or activities—called foci. Here, the foci are replaced by values, and links representing relationships between values are added as well. This kind of mapping graph could be of great use for the conceptualization of the whole system.

It is worth noting that, affiliation networks, if combined with social network, emphasize the relationships between people, but the graph shown here doesn't show the links between stakeholders, because such relationships are not of the great importance of value system. Relationships between stakeholders, however, could influence value system as well, especially within an organization. For example, the values of the management department of the mall could impact that of the sellers, causing a "triadic closure." So, designers should keep in mind that the power structure of an organization could influence value system, and this kind of dynamic process could be explored in the empirical investigation as well.

Empirical Investigation. Based on the stakeholders and values identified in conceptual investigations, designers may empirically explore how the stakeholders interact with the system in a potential usage scenario and how different designs might result in different outcomes in value-building. There are a lot of methods to choose from. For example, one VSD case study in *The Handbook of Information and Computer Ethics* that concerns how high-definition plasma displays in an interior office could benefit employees psychologically used multiple empirical methods, including physiological data analysis (heart rate), behavior data analysis (eye gaze), interviews, and surveys.[61] There are also many novel methods that are applicable, for example, web data analysis using web crawlers

and natural language processing (NLP), agent-based modeling (ABM), and social network analysis (SNA). In short, any empirical methods that can help understand how human values interact with the technology can be employed.

Since FRTs with social sorting functions are likely to be used in discriminatory applications, at least one empirical investigation related to discrimination must be conducted. For example, for a shopping recommendation system that uses an FRT with gay-recognition functions to assist sellers to decide what kinds of goods a shopper at a mall might prefer, at least one empirical study must be carried out to collect information such as how people with various sexual orientations in both customers and sellers see this function, how this function affects their decision-making in buying and selling, how their psychological well-being would be influenced, whether they feel discriminated against, whether they feel their privacy is being violated, and how they prioritize equality and efficiency, etc. This empirical investigation can both provide insight to further perfect the conceptualized value system of the technology and help decide whether and how a property of the technology should be designed. For example, if the empirical investigation shows that the financial benefit brought about by gay-recognition function in the recommendation FRT would fail to compensate for the harms to value-building, then the engineers should think twice on whether this function should be developed and integrated into the system in the first place.

Technical Investigation. Based on the conclusions drawn from the previous two set of investigations, designers could be able to make designing decisions that align with the value system. In the case of FRT, the decisions include whether and how certain FRT function should be integrated into a system, for example, whether a gay-recognition

function should be built into an airport security system. If the answer is no, how can such pre-existing biases be removed? If the answer is yes, how could the component be designed in a way that minimizes discrimination? For example, if an FRT system has an obvious racial bias against African Americans that could lead to discrimination, technical approaches should be explored to eliminate or compensate such biases. It might not be easy, from an engineering point of view. The most intuitive solution would be adding photos of less-represented race into the training data. For example, Modiface, a Toronto-based FRT company pays for extra images in order to enhance its database of ethnic minority groups.[67] However, according to Jonathan Frankle, a former staff technologist for the Georgetown University Law Center, simply adding more photos of African Americans into the training data will not solve the problem perfectly. "If it were just about putting more black people in a training set, it would be a very easy fix. But it's inherently more complicated than that," as he put it, since other factors matter as well, such as the relatively more difficult technical challenge of pinpointing landmarks on darker skin.[68] There are people trying to solve the problem. For example, an FRT startup Gfycat's FRT was found to have troubles recognizing Asians because there are fewer Asian faces in the training dataset, so they worked out a solution of building a kind of Asian-detector. When an Asian photo is fed into the system, a more sensitive mode will be turned on so that the threshold of matching will become stricter. It is claimed that through this approach, the accuracy for Asians has been improved to 93%, while the accuracy for white people is 98%.[67] However, there are also some people that think that AI systems shouldn't profile race.[69]

2. The limitations of IRB in the design of Facial Recognition Technology research

Some suggestions for designing FRT products were given in the previous section. What about FRT research? What principles should be included in the design and implementation of FRT research? And what implications do they have to FRT practices?

As mentioned in Chapter 3, IRB aims to protect human subjects in research and make sure research being "carried out in an ethical manner." The basic ethical principles of IRB include respect for persons, beneficence, and justice. However, IRB is doing poorly with data science and AI. The Stanford Gay face recognition research was approved by IRB, but it doesn't mean the research is completely ethical. A lot of things need to be addressed.

One important requirement of IRB is for the informed consent of the "subject or the subject's legally authorized representative." Informed consent should provide each subject with information such as the purposes of the research, risks or discomforts, benefits, alternative procedures, and contact information. But exceptions exist. According to the Federal Policy for the Protection of Human Subjects, or the "Common Rule," which outlines the basic provisions of IRB to regulate biomedical and behavioral research involving human subjects, there are some situations where the requirements of informed consent can be waived, when:

"*(1) The research involves no more than minimal risk to the subjects; (2) The waiver or alteration will not adversely affect the rights and welfare of the subjects; (3) The research could not practicably be carried out without the waiver or alteration; and (4) Whenever appropriate, the subjects will be provided with additional pertinent information*

after participation."[70]

Does FRT research satisfy the conditions above? The minimal risk exception means the harm or discomfort that could be brought about by the research is smaller than those "ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests." In this sense, it seems most FRT research brings subjects minimal risks since there's no physical contact, no human encounter, and no direct consequences. Many face photos used in FRT research are collected from the internet, for instance, through social media or dating apps. It also seems not practical to ask for approval from each subject, since the number is astronomically huge.[71] But FRT research brings many new issues that are not fully realized in traditional biomedical research, such as the potential of developing discriminatory applications, as we mentioned in previous chapters. For example, if someone whose photos in dating apps are collected in the training data set for Stanford gay recognition research is aware of the existence and the potential discriminatory applications of the research, he/she might not be OK with it. Does he/she have the right to know about this research? Can he/she ask them to remove his/her photos from the database? If he/she finds the truth after the training of the FRT algorithm is done, how could he/she remove the training results that come out of his/her data, since simply deleting the photos is of no use? Those issues could not be answered by IRB as used today.

IRB has many limitations when applied in data science in general, since the original purposes of IRB limit its applicability in data science. IRB was initially designed to protect human subjects in biomedical and behavioral research, because before IRB, abuses of human subjects in these kinds of research drew much public attention, especially during the WWII, such as the Tuskegee Syphilis Study and Stanford prison experiment. Now,

IRBs are widely used in the areas of healthcare, sociology, and psychology, focusing mainly on how the research affects human subjects. As mentioned in the previous paragraph, research in data science involves much less engagement of human subjects. This is shown in two aspects.

First, fewer human subjects are participating in data science. Most data science research, especially AI research, try to find new insights into existing data, instead of creating new data from scratch. This decreases the need for methodologies involving human subjects such as interviews, surveys, focus groups, and experiments, which data scientists seldom do. Most of the time, they just take data from previous records.

Second, the way human subjects participate in data science research is non-traditional and indirect, making it hard to pinpoint their participation and predict the potential harm. According to the Common Rule, human subject means "a living individual about whom an investigator conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information."[70] Traditional research normally involves both of them. Data science, however, is more ambiguous. Data science seldom collects data through intervention, as discussed above. The tricky part is whether data science, especially AI research, obtain "identifiably private information." According to the Common Rule, the identifiable private information includes information that an individual can "reasonably expect" will not be "made public." For example, an employee may complain about a company policies in an interview of an organization communication research. The details of the interview, even without personal information, could be used to identify the interviewee's identity, which might cause him/her trouble in the workplace. Any records or transcripts of the interview is a piece of identifiable private

information. But data scientists seldom obtain identifiable private information by themselves. With very few exceptions, they usually use previously obtained private information as data to produce new insights and draw novel conclusions. For example, Stanford gay recognition research didn't take photos and collect information such as sexual orientation by themselves. Instead, they use "self-taken images obtained from online dating websites," just as a bunch of other FRT research, as mentioned in their paper. The reason, as they explained, is because in this way, "images can be collected in large numbers, from more representative samples, and at a lower cost." [2] Therefore, in IRB's standards, the personal information used in FRT research is neither about behavior nor provided by the individuals, so there is space for not being seen as "identifiably private information." However, this information *is* literally about behavior (sexual orientation is nothing less about behavior), and *is* identifiable and private! The only reason that the information is not provided by the individuals is they don't even know their photos are used in such research. So, the "human subject" definitions of IRB are somehow outdated and should be modified for data science.

Moreover, IRB only reviews the well-being of human subjects, but nothing about the fundamental ethics about the research in question and whether it should be carried out in the first place. This is more like a focus on the details instead of the whole picture. Consider the Stanford gay recognition research, which was approved by IRB, it is hardly possible to include potential discriminatory applications into the consideration of a standard IRB review, since those applications would not directly affect the welfare of the people in the training set, but they surely might have some non-negligible consequences. Risks do not necessarily rest in the treatment of human subjects. They could rest in the

purposes, design, conduct, management, and consequences embedded in the socio-technical system of research. An analogous example that could provide some insight is nuclear ethics, which not only considers the welfare of people directly associated with nuclear research like human radiation experiments, but also a wide range of other issues, such as environmental problems of uranium mining, nuclear accidents, labor, and freedom of speech. The principles of nuclear ethics include righteous defense, minimizing nuclear harm, eliminating the risk of nuclear wars, and ensuring the world peace,[72] though debates are still going on. Today, just like nuclear technology, AI is seen by some people as a new kind of WMD—weapon of "Math" destruction.[73] It becomes a moral imperative to develop similar ethical principles to evaluate the potential risks of FRT and AI in general.

Given the fact that IRB might not be completely applicable for data-driven research, in August 2012, the Menlo Report was released by the U.S. Department of Homeland Security Science & Technology Directorate, Cyber Security Division, to serve as a guideline for research involving Information and Communications Technologies (ICT). The three principles of Belmont Report—respect for persons, beneficence, and justice were adapted into Menlo Report, and a fourth principle was added—respect for law and public interest, which states "engage in legal due diligence; be transparent in methods and results; be accountable for actions." In the justice principle, Menlo Report emphasized that the initial selection of the subjects should be guided by fairness, and research "should not arbitrarily target persons or groups based on attributes including (but not limited to): religion, political affiliation, sexual orientation, health, age, technical competency, national origin, race, or socioeconomic status." This principle can serve as a guideline for the selection of FRT training data. Also, the fourth principle underlines the "transparency of

methodologies and results, and accountability of actions," which, in FRT context, can be interpreted as the transparency of the training data, the biases of the algorithms, and how the study is actually carried out, so that the general public can assess the whole process and the risks related to the research.

3. Beyond boundaries: A case study

In June 2014, Facebook, the biggest social media in the world, published a research paper on *Proceedings of the National Academy of Science*. In the study, they altered the feeds presented to 689,003 users to see whether the emotion of a user's posts would be altered if he/she exposes to certain emotional content.[74] This paper sparked widespread outrage due to the ethical issues of emotional manipulation and was criticized as the "emotional contagion" study. Debates centered around whether it is ethical to manipulate people's emotion without informed consent. But the truth is, this kind of manipulation is the basis of almost all products of all social media platforms.[75]

The case of Facebook's research is a perfect example where the boundary between academic research and industrial practice is blurred. For one thing, it was carried out in a for-profit organization, with the potential to improve the advertisement system and recommendation system in order to create more lucrative products. For another, it could help understand human psychology and behavior, with high academic values. The discussions around this example could also provide profound insight and implication to FRT industrial research.

Most of the critiques of the "emotional contagion" study are to question whether it undermines people's emotional well-being, why it is unethical, and how to improve future

studies. Some people think IRBs or similar ethical-reviewing procedures should be required before research at industrial organizations being carried out so that ethical issues could be mitigated, just like the academia does. For example, some legal scholars propose doing "consumer subject review boards" in industrial research.[76] However, as Microsoft researcher Danah Boyd points out, adopting IRB would not necessarily make Facebook and other for-profit companies more ethical. Just like the Stanford gay recognition FRT, many controversial studies, including the "emotion contagion" study, would "likely pass an IRB examination," as Boyd put it. Besides the downsides of IRB in data science discussed in previous sections, IRB is also criticized as being inconsistent among institutes. Moreover, ethicists have expressed concerns about the ethical value of IRB at all, even for academic research.[77] "Ethics aren't a checklist. Nor are they a universal," as Boyd put it. He thinks ethics should be integrated into everyday practices, as opposed to being "outsourced" to external review. To achieve that goal, researchers and practitioners should be trained on ethics, so that they could better understand ethical aspects of their research and products, as well as incorporate this knowledge into their daily practice. Therefore, as he proposed, a "socio-technical model of ethical oversight" should be constructed by the companies and researchers who are involved and who truly understand the process of R&D and the decision-making process of the company, instead of some organizations from outside.

Facing the pressure from the public and scholars, Facebook quickly came up with an ethics evaluation procedure and their public policy researchers Molly Jackman and Lauri Kanerva published a paper explaining what the system does and how it works. They first reviewed the limitations of IRB in providing context-specific guidelines for industrial

research, especially big data research and emphasized that "there is no one-size-fits-all" model. Then they outlined the procedure developed to review future research, partially in order to avoid repeating the "emotion contagion" mistake. Basically, the procedure consists of three main parts—training, which includes three levels—employee onboarding, researcher-specific training, and reviewer-specific training, review by substantive area experts, and review by research review groups, as the figure shows below. Four criteria were compiled to evaluate the ethics—how the research will improve the society, whether there are "potentially adverse consequences," whether it is consistent with people's expectations, and whether there is proper protection of personal information.[78]

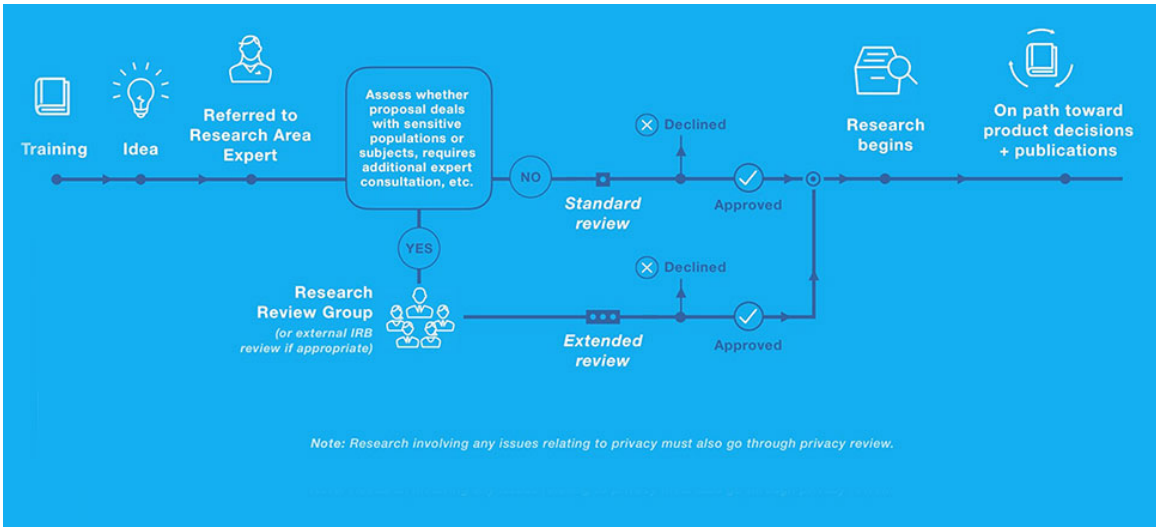

Figure 2 Research Review at Facebook (Jackman & Kanerva, 2016)

Facebook's procedure in some sense conforms with Boyd's viewpoint. On the one hand, training on ethics is emphasized as the first step of the process; on the other hand, it considers a variety of factors and intends to build an inclusive reviewing networks from multiple stakeholders, including internal experts like the senior managers of the research teams, as well as external experts from law, ethics, communication, and policy sectors, so that the discussion could be built on common ground of multiple disciplines. In particular,

as a demonstration, they collected feedbacks from LGBT groups before studies on LGBT trend on Facebook were carried out, in order to make sure it was ethical and perspectives from various stakeholders were included into the consideration.

Facebook's system is a good example of industrial research. It is quite different from IRB in terms of procedures and how the review teams are constructed, though it does share a basic formula with IRB— "considering the benefits of the research against the potential downsides." It has many implications for industrial research related to FRT. Basically, two main changes could be initiated. First of all, FRT companies may train employees on ethics, including the ethical principles in general and the issues of discrimination in particular. Like Facebook, the training could be done in several levels, including general training for all employees, specific training for researchers, and training for reviewers. VSD, which was mentioned in the previous sections, may be of substantial help if included in the training. Second, one or more review groups and reviewing procedures should be constructed for reviewing research proposal beforehand, overseeing research during the process, estimating the ethics of the research results, and dealing with potential ethical discussion or controversies afterward. Particularly, external consultation with organizations representing potential discriminatees could be very helpful. For example, for a Google FRT application that is used to categorize photo album in cell phones, if the accuracy varies a lot among ethnic groups, for example, mistakenly labelling African Americans as "Gorillas" in the beta version, then consulting with organizations like National Association for the Advancement of Colored People (NAACP) and even including them into the reviewing process before the release of official version would significantly reduce the ethical risk of discrimination.

Facebook's system, however, is not universally helpful, since Facebook is a top tech giant with enormous resources to spend. Actually, the system Facebook designed is built on their existing infrastructure. For example, the review process is managed on Facebook's internal online task-tracking platform, so that the process could be easily integrated into researchers' everyday workflow. Also, they have extensive external networks with government agencies, commercial companies, and other organizations. Not all FRT companies have such a strong resource to work on, and many of them are just startups. But it does not mean ethics review process is unnecessary for them. There are many avenues for adjustment according to the actual situations. For example, industry associations and NGOs could provide some general ethical guidelines, as well as consulting services regarding legal issues, training, public relations, crisis communications, and other services. Also, industry associations can work with individual companies to help them construct review boards, build networks, negotiate with external agencies, and increase bargain power collectively. But the process should not be outsourced entirely to external organizations or review boards like IRB, because the employees of the company, be them researchers or managers, are the people who know the company best in every aspect, so their engagement is the key to the success of the process. For example, ethicists and legal experts may argue there are serious ethical issues associated with an FRT that identifies white males better than black females. Their argument is definitely true, but they could not provide insights or advice on how to improve the algorithm since it's a complicated, multifactor, socio-technical problem. As we discussed in previous chapters, simply increasing the percentage of black females in the training dataset cannot perfectly solve the problem. Whatever the solutions may be, the engineers of the FRT company are the people

who know their products best, so they know how to mitigate the problems with minimal costs, and the managers have the knowledge of how to facilitate the product iteration process and how to seek supports and endorsements from the executives of the company.

CHAPTER 5

CONCLUSION AND DISCUSSION

1. Conclusion

FRT stems from some early ideas that the process of recognizing people's identities could be automated and that some features on human bodies in general and human faces, in particular, could indicate their characteristics and personalities. The latter idea stems from physiognomy that has long been refuted, but it is scientifically evident that links between facial features and personalities exist.

The earliest pioneers of FRT in the 19th century tried to compile a table by measuring facial features manually. Since the introduction of digital computers in the 20th century, significant progress has been made in this area. The early FRTs were semi-automated, with facial landmarks manually coded into computers. In the late 1980s, a milestone system called Eigenfaces, which identified faces through their deviations from the average, improved the efficiency of FRT significantly. In recent years, with the integration of advanced AI techniques and pattern recognition algorithms, FRT gained considerable momentum and the accuracy soon skyrocketed to the point that it surpassed human capacity. This was realized by using significant amounts of face images to train FRT algorithms. Also, FRT is increasingly incorporated into existing systems, such as CCTV systems and smart devices, and new applications of real-time FRT and 3D FRT are currently under development. Basically, the process of current AI-powered FRT consists of face detection, facial features extraction to form a "faceprint," and matching in target databases or classifying by categories.

With the highly improved accuracy and great cost efficiency, FRT has enormous potentials. Nowadays, FRT is used in a wide range of areas, such as surveillance, banking, e-commerce, and games. At the same time, a lot of researchers are doing FRT-related studies, trying to improve the algorithms or use FRT to test hypotheses in other areas. For example, researchers at Stanford University published a study that used FRT to predict people's sexual orientation based on their facial features, and this study, according to the author, tested a widely accepted hypothesis called prenatal hormone theory (PHT), which associates facial appearances with sexual orientations with fetal androgen signaling on sexual differentiation. The Stanford gay recognition algorithm is claimed to be more accurate than human predictions since it could identify very subtle features that are unrecognizable to the naked eyes.

Recently, however, FRT has aroused considerable ethical controversy. For example, the Stanford gay recognition FRT sparked widespread outrage among LGBTQ communities since it could be used to develop discriminatory applications. The FRT used in the 2001 Super Bowl in Tampa, Florida, which snapped every spectator's face and matched against a database of criminals, raised considerable privacy concerns. The criminal prediction FRT developed by researchers at Shanghai Jiaotong University in China raised concerns of totalitarian scenarios.

The applications of FRT can be categorized into two types—recognition and classification. Recognition applications focus on matching captured face images and searching for a match in databases, while classification applications don't actually identify people but classify people according to attributes such as gender, race, and sexual orientation. Recognition applications are mainly associated with ethical issues such as

privacy, while classification tasks may cause concerns about discrimination. This thesis mostly focuses on the latter.

There are three main ethical frameworks that can be used to analyze the ethical questions associated with technologies—utilitarianism, deontology, and virtue theory. Utilitarianism evaluates the consequences of actions, with the aim of creating "the greatest good for the greatest number." Deontologists like Kant argue that the ethical standard of behavior is to conform with a maxim that should be "unquestionably universalized." They also emphasize the importance of respecting people as autonomous agents, and oppose treating people as "mere means." Virtue theory, which could be traced back to Plato and Aristotle, focuses on the characters of the actors, for instance, honesty, patience, courage, and other traits that can promote excellence. The three frameworks have been woven into numerous social practices, and IRB is one of them. IRB is a review procedure in academic research in order to protect human subjects. The basic ethical principles of IRB are respect for persons, beneficence, and justice. However, IRB is not the panacea to address ethical issues associated with FRT, because it has limitations in data science. It only emphasizes protecting human subjects, and it is not required in industrial research. The controversial Stanford gay recognition study did pass the IRB review.

Discrimination is the main issue associated with the classification applications of FRT. Why is discrimination wrong? Theoretically, discrimination is treating people differently based on the differences in some traits, such as gender, age, and race. Some types of discrimination raise no obvious ethical issues, for example, the discount given to Georgetown students in the stores on M Street, while some others do, when the differentiated treatments are based on the "membership of a socially salient group," such

as gender, religion, and sexual orientation. This type of discrimination is morally wrong because, from a utilitarian point of view, it undermines the ability of individuals and the society as a whole in creating "the greatest good for the greatest number." Also, from a deontologist point of view, discriminatory actions treat people unfairly on the basis of irrelevant characteristics and awfully undermine social equality. Furthermore, from a virtue theory point of view, discriminatory behaviors hurt the virtuous characters on the level of both individual and community.

FRT applications with classification functions have high risks of causing discrimination. First of all, FRT is inherently biased due to the training data. For example, the FRT algorithm used by FBI is mainly trained on white males, so it is less accurate on females and people of colors. It might cause serious problems since this algorithm is actually deployed in law enforcement. Second, this kind of FRT could be easily used to develop applications that automate discriminatory actions. For example, the Stanford gay recognition algorithm could be easily integrated into existing systems and unwittingly treat homosexual people differently. Third, FRT with classification functions may increase the number of people who suffer from discrimination in two ways, one of which is that it may add new groups of discriminatees, while the other is that it may increase the number of people in each discriminated group. Moreover, discrimination of FRT is difficult to discover and mitigate since the algorithms are mainly black-boxed.

What can be done to mitigate the discrimination issues of FRT? To address this question, ethical considerations must be integrated into the design process of FRT. Three types of practice in FRT can be identified—industrial product development, academic research, and industrial research.

For industrial product development, Value Sensitive Design (VSD) is a very useful approach to incorporate human values into the design. The methodology of VSD consists of three categories of investigations—conceptual, empirical, and technical. In conceptual investigations, the direct and indirect stakeholders of the FRT application must be identified. Particularly, the algorithm must be tested beforehand to find out any possible biases, and the groups of people that are biased against must be included into stakeholders. For example, if the Stanford gay recognition FRT were to be used in a commercial application, then the homosexual clients must be listed as a direct stakeholder. Then, the key values of each group of stakeholders must be identified. I propose using a "value network" to visualize the relationships between stakeholders and values. If the different groups of stakeholders are in the same organization, the power structure of the organization is also important. Based on the stakeholders and their values, empirical investigations use a wide range of methods from social science, psychology, computer science, and other areas to empirically explore how the stakeholders interact with the FRT product under design. For FRT with classification functions, at least one empirical investigation related to discrimination issues must be conducted. In the technical investigation, the value conflicts and ethical issues identified in the two previous investigations are the central considerations. For example, for an FRT that is less accurate when working on people of color, some technical methods must be employed to explore how to mitigate the issues, settle value conflicts, and balance the interests of different stakeholders, for example, by putting more photos of ethnic minority groups into the training data, or building more sensitive detectors or increasing the detecting threshold for people with darker skins.

For academic research related to FRT, the common procedure to evaluate the ethics

beforehand is IRB, but IRB was designed to protect human subjects mainly in biomedical and behavioral research. It has many limitations in the new context of data science and machine learning since data science involves much less engagement of human subjects. For one thing, fewer human subjects are participating in data science research; for the other, the way human subjects participate in data science research is indirect, so it is hard to assess their participation and predict the potential harm. Moreover, IRB only considers the well-being of human subjects, but nothing to assess whether such study should be carried out in the first place. The Menlo Report, which was released in 2012 by U.S. Department of Homeland Security, adapted the principles of IRB and added the fourth principle, extending its applicability for ICT research. It provides some insights on how to assess and regulate FRT research.

For industrial research, which is conducted in companies, as opposed to in academia, the situation is more complicated. IRB is only required for academic research that involves human subjects, so it is not applicable for industrial research. A good example of industrial research and how to deal with its ethical issues is the notorious "emotion contagion" study published by Facebook in 2012. This study received tons of ethical critiques. Facebook quickly came up with an ethics reviewing system, providing an ethical guideline for future studies. They employed some principles from IRB, but build the system on the basis of its internal infrastructure. The system emphasizes the importance of ethical training and seeking help from external experts. As Microsoft researcher Danah Boyd points out, ethical training, instead of outsourcing the process to external review boards, could better help integrate ethics into the company's everyday practice. Facebook's reviewing system, however, is not universally helpful, since Facebook is one of the top tech giants who have

a huge amount of internal resources, existing infrastructures, and extensive external networks. Industry associations and NGOs related to FRT might be helpful in providing ethical guidelines and services on legal issues, training, public relations, and other issues.

2. Further discussion

Other ethical issues

Though this thesis mainly focuses on discrimination issues in FRT, as discussed before, FRT also raises ethical concerns such as privacy, autonomy, and surveillance. The design thinking in Chapter 4 can be adapted to mitigate those issue as well.

Classification FRT could cause issues besides discrimination. For example, privacy is another concern. In the case of the Stanford gay recognition FRT, there are people who don't want to disclose their sexual orientation, so the utilization of this FRT certainly has a risk of violating people's privacy. According to the discussion in Chapter 4, further assessment other than IRB should be conducted to see if this research should be carried out in the first place. Also, using VSD, companies may determine if the function should be incorporated into the system, how people perceive the privacy related to this function, and how to incorporate it so that the violation of privacy could be minimized.

In particular, for privacy issues, which are currently in hot debate, an approach called privacy by design was developed to build privacy concerns in the initial design of data system so that the data is protected from the very beginning, as opposed to seeking solutions after data is hacked.

Identification FRT used in surveillance may cause trust and autonomy problems. It limits the autonomy of individuals by limiting one's options of action. If a person knows

someone is watching him or her, even in a non-obvious way, he/she will change his/her behavior according to some norms, instead of choosing it from free will. Also, it might cause chilling effects for the right of free speech.

Other design methods and principles

In Chapter 4, we talked about how to design less discriminatory FRT systems in three different contexts. There are also other methods that can be used as supplements to help identify problems and build human values into the systems. One example is the social-systems analysis proposed by Kate Crawford and Ryan Calo, which "engages with social impacts at every stage—conception, design, deployment, and regulation."[79] A social-systems approach looks at the problems from multifaceted angles such as sociology, anthropology, science and technology studies (STS), law, and philosophy. It requires researchers to explore how different communities' employment of data and resources could affect AI algorithms that are trained on those data. This could be extremely helpful in interpreting anomalies and identifying roots of issues. For example, in a 2015 study, a machine learning algorithm used in a hospital made a serious mistake that instructed doctors to send patients with asthma home because historically those people were automatically treated with "intensive care" hence no "required further care" records could be found in the database on which the algorithm was trained. Through social-systems analysis, the underlying logic of the result could be examined. This approach can be helpful in FRT R&D as well.

Also, in recent years, with the increasing awareness of both practitioners and the general public, many conferences and workshops have been held in order to discuss the principles of machine learning and AI in general. For example, Fairness, Accountability,

and Transparency in Machine Learning (FAT/ML) is an annual event where researchers explore how to address the issues such as discrimination. They identified five principles for accountability: responsibility, explainability, accuracy, auditability, and fairness.[80]

Another example that drew much attention was the Asilomar Conference on Beneficial AI organized by the Future of Life Institute (FLI) in January 2017 at Asilomar, California. The participants include over one hundred experts and researchers from various areas such as philosophy, computer science, law, ethics, and economics. With the goal of creating AI that is beneficial to humanity, *Asilomar AI Principles* was generated from the conference. The principles consist of three main parts—research issues, ethics and values, and longer-term issues, and gave twenty-three general principles for governments, AI researchers, and practitioners as guidelines. Some of the principles, especially in the Ethics and Value part, are very relevant to the discrimination issues of FRT and deserve serious consideration, for example, the principles of Failure Transparency, Judicial Transparency, Responsibility, and Human Value.[81]

Another example that could provide insights to FRT industrial research is the ethical reviewing procedure compiled by the Information Accountability Foundation (IAF) for industrial research related to big data. This document results from one of IAF's projects called Big Data Ethics Initiative. The ethical framework consists of five key values—beneficial, progressive, sustainable, respectful, and fair.[82] Based on the ethical principles, IAF gives a comprehensive assessment worksheet, which estimates the aspects like purpose, sources, preparation, and stakeholders of the project, as well as the five ethical principles. This could also help FRT companies to compile similar frame and worksheet according to their actual situation.

For future research, empirical studies could be conducted to see how FRT actually interact with stakeholders and how it affects human values. Also, for FRT practitioners, it could be helpful to compare how different techniques reduce biases.

BIBLIOGRAPHY

[1] "Advances in AI Are Used to Spot Signs of Sexuality," *The Economist*, September 9, 2017, https://www.economist.com/news/science-and-technology/21728614-machines-read-faces-are-coming-advances-ai-are-used-spot-signs.

[2] Yilun Wang and Michal Kosinski, "Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images.," *Open Science Framework*, February 15, 2017, https://doi.org/None.

[3] Sam Levin, "LGBT Groups Denounce 'dangerous' AI That Uses Your Face to Guess Sexuality," the Guardian, September 9, 2017, http://www.theguardian.com/world/2017/sep/08/ai-gay-gaydar-algorithm-facial-recognition-criticism-stanford.

[4] Xiaolin Wu and Xi Zhang, "Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of ArXiv:1611.04135)," *ArXiv:1611.04135 [Cs]*, November 13, 2016, http://arxiv.org/abs/1611.04135.

[5] Emerging Technology from the arXiv, "A Deep-Learning Machine Was Trained to Spot Criminals by Looking at Mugshots," MIT Technology Review, accessed January 19, 2018, https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/.

[6] "Firm Defends 'snooper Bowl' Technology," CNET, accessed January 19, 2018, https://www.cnet.com/news/firm-defends-snooper-bowl-technology/.

[7] "Is America Really the Land of the Free?," Text.Article, Fox News, July 9, 2001, http://www.foxnews.com/story/2001/07/09/is-america-really-land-free.html.

[8] S. A. Mathieson, "In Sight of the Law," the Guardian, March 1, 2001, http://www.theguardian.com/technology/2001/mar/01/onlinesupplement.

[9] Alexandra Stikeman, "Recognizing the Enemy," MIT Technology Review, accessed January 20, 2018, https://www.technologyreview.com/s/401300/recognizing-the-enemy/.

[10] "Facial Recognition May Boost Airport Security But Raises Privacy Worries," NPR.org, accessed January 20, 2018, https://www.npr.org/sections/alltechconsidered/2017/06/26/534131967/facial-recognition-may-boost-airport-security-but-raises-privacy-worries.

[11] Clare Garvie, *The Perpetual Line-up: Unregulated Police Face Recognition in America* (Washington, DC: Georgetown Law, Center on Privacy & Technology,

2016).

[12] A. J. O'Toole et al., "Face Recognition Algorithms Surpass Humans Matching Faces Over Changes in Illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, no. 9 (September 2007): 1642–46, https://doi.org/10.1109/TPAMI.2007.1107.

[13] B. F. Klare et al., "Face Recognition Performance: Role of Demographic Information," *IEEE Transactions on Information Forensics and Security* 7, no. 6 (December 2012): 1789–1801, https://doi.org/10.1109/TIFS.2012.2214212.

[14] Joy Buolamwini, "Media Lab Student Wins National Award for Fighting Bias in Machine Learning," MIT Media Lab, accessed February 2, 2018, https://www.media.mit.edu/posts/media-lab-student-recognized-for-fighting-bias-in-machine-learning/.

[15] Maggie Zhang, "Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software," Forbes, accessed February 2, 2018, https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/.

[16] Helen Nissenbaum, "Privacy as Contextual Integrity Symposium - Technology, Values, and the Justice System," *Washington Law Review* 79 (2004): 119–58.

[17] Philip Brey, "Ethical Aspects of Facial Recognition Systems in Public Places," *Journal of Information, Communication & Ethics in Society* 2, no. 2 (2004): 97–109.

[18] Alfred C. Kinsey, Wardell Baxter Pomeroy, and Clyde Eugene Martin, *Sexual Behavior in the Human Male* (Philadelphia: W. B. Saunders Co, 1948).

[19] Louis A. Knafla, *Policing and War in Europe* (Greenwood Publishing Group, 2002).

[20] George Pavlich, "The Subjects of Criminal Identification," *Punishment & Society* 11, no. 2 (April 1, 2009): 171–90, https://doi.org/10.1177/1462474508101491.

[21] W. Zhao et al., "Face Recognition: A Literature Survey," *ACM Computing Surveys (CSUR)* 35, no. 4 (January 12, 2003): 399–458, https://doi.org/10.1145/954339.954342.

[22] A. J. Goldstein, L. D. Harmon, and A. B. Lesk, "Identification of Human Faces," *Proceedings of the IEEE* 59, no. 5 (May 1971): 748–60, https://doi.org/10.1109/PROC.1971.8254.

[23] L. Sirovich and M. Kirby, "Low-Dimensional Procedure for the Characterization of Human Faces," *JOSA A* 4, no. 3 (March 1, 1987): 519–24,

https://doi.org/10.1364/JOSAA.4.000519.

[24] Matthew Turk and Alex Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience* 3, no. 1 (January 1, 1991): 71–86, https://doi.org/10.1162/jocn.1991.3.1.71.

[25] P. Viola and M. Jones, "Robust Real-Time Face Detection," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, 747–747, https://doi.org/10.1109/ICCV.2001.937709.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *ArXiv:1503.03832 [Cs]*, June 2015, 815–23, https://doi.org/10.1109/CVPR.2015.7298682.

[27] Jon Russell, "China's CCTV Surveillance Network Took Just 7 Minutes to Capture BBC Reporter," *TechCrunch* (blog), accessed January 30, 2018, http://social.techcrunch.com/2017/12/13/china-cctv-bbc-reporter/.

[28] Javier C. Hernández, "China's High-Tech Tool to Fight Toilet Paper Bandits," *The New York Times*, March 20, 2017, sec. Asia Pacific, https://www.nytimes.com/2017/03/20/world/asia/china-toilet-paper-theft.html.

[29] "Coke Trials Facial Recognition Vending Machines in Australia," *BiometricUpdate* (blog), June 3, 2014, http://www.biometricupdate.com/201406/coke-trials-facial-recognition-vending-machines-in-australia.

[30] Carmen E. Lefevre et al., "Telling Facial Metrics: Facial Width Is Associated with Testosterone Levels in Men," *Evolution and Human Behavior* 34, no. 4 (July 1, 2013): 273–79, https://doi.org/10.1016/j.evolhumbehav.2013.03.005.

[31] Benedict C. Jones et al., "Facial Coloration Tracks Changes in Women's Estradiol," *Psychoneuroendocrinology* 56, no. 10.1016/j.psyneuen.2015.02.021 (June 2015): 29–34, https://doi.org/10.1016/j.psyneuen.2015.02.021.

[32] Karel Kleisner, Veronika Chvátalová, and Jaroslav Flegr, "Perceived Intelligence Is Associated with Measured Intelligence in Men but Not Women," *PLOS ONE* 9, no. 3 (March 20, 2014): e81237, https://doi.org/10.1371/journal.pone.0081237.

[33] Mare Lõhmus, L. Fredrik Sundström, and Mats Björklund, "Dress for Success: Even Unseen Clothing Increases Female Facial Attractiveness," *Annales Zoologici Fennici* 46 (February 27, 2009): 75–80.

[34] Malvina N. Skorska et al., "Facial Structure Predicts Sexual Orientation in Both Men and Women," *Archives of Sexual Behavior* 44, no. 5 (July 1, 2015): 1377–94, https://doi.org/10.1007/s10508-014-0454-4.

[35] Toan Thanh Do and Thai Hoang Le, "Facial Feature Extraction Using Geometric Feature and Independent Component Analysis," in *Knowledge Acquisition: Approaches, Algorithms and Applications*, Lecture Notes in Computer Science (Pacific Rim Knowledge Acquisition Workshop, Springer, Berlin, Heidelberg, 2008), 231–41, https://doi.org/10.1007/978-3-642-01715-5_20.

[36] Y. Weiwei and Y. Nannan, "Facial Feature Extraction on Fiducial Points and Used in Face Recognition," in *2009 Third International Symposium on Intelligent Information Technology Application*, vol. 3, 2009, 274–77, https://doi.org/10.1109/IITA.2009.241.

[37] Hua Gu, Guangda Su, and Cheng Du, "Feature Points Extraction from Faces" (Image and Vision Computing, New Zealand, 2003), 154–58, http://sprg.massey.ac.nz/ivcnz/Proceedings/IVCNZ_28.pdf.

[38] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, no. 6 (June 2001): 681–85, https://doi.org/10.1109/34.927467.

[39] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep Face Recognition," vol. 1, 2015, 6.

[40] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Transactions on Circuits and Systems for Video Technology* 14, no. 1 (January 2004): 4–20, https://doi.org/10.1109/TCSVT.2003.818349.

[41] Alan K. L. Chan, ed., *Mencius: Contexts and Interpretations* (University of Hawai'i Press, 2002), http://www.jstor.org.proxy.library.georgetown.edu/stable/j.ctt6wr328.

[42] Laozi, Takuan Sōhō, and Thomas F. Cleary, *Tao Te Ching: Zen Teachings on the Taoist Classic*, 1st ed (Boston : [New York]: Shambhala ; Distributed in the U.S. by Random House, 2010).

[43] "Applied Ethics - Philosophy - Oxford Bibliographies - Obo," accessed February 4, 2018, http://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0006.xml.

[44] Louis P. Pojman and Peter Tramel, eds., *Moral Philosophy: A Reader*, 4th ed (Indianapolis: Hackett, 2009).

[45] Joseph Priestley, *An Essay On the First Principles of Government: And On the Nature of Political, Civil, and Religious Liberty* (Nabu Press, 2010).

[46] John Stuart Mill and Colin Heydt, *Utilitarianism*, Broadview Editions (Peterborough, Ont: Broadview Press, 2011).

[47] Immanuel Kant and H. J. Paton, *Groundwork of the Metaphysic of Morals*, 1st Harper Torchbook ed, Harper Torchbooks ; TB 1159 (New York: Harper & Row, 1964).

[48] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, "Belmont Report: Ethical Principles And Guidelines For The Protection Of Human Subjects Of Research (1979)," Text, HHS.gov, January 28, 2010, https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html.

[49] Michel Foucault, *Discipline and Punish: The Birth of the Prison*, 2nd Vintage Books ed (New York: Vintage Books, 1995).

[50] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, "Belmont Report: Ethical Principles And Guidelines For The Protection Of Human Subjects Of Research (1979)."

[51] W. D. Ross and Philip Stratton-Lake, *The Right and the Good*, New ed. (Oxford : New York: Clarendon Press ; Oxford University Press, 2002).

[52] "AI Research Is in Desperate Need of an Ethical Watchdog," WIRED, accessed February 6, 2018, https://www.wired.com/story/ai-research-is-in-desperate-need-of-an-ethical-watchdog/.

[53] "Discrimination Meaning in the Cambridge English Dictionary," accessed March 4, 2018, https://dictionary.cambridge.org/dictionary/english/discrimination.

[54] Kasper Lippert-Rasmussen, *The Routledge Handbook of the Ethics of Discrimination*, Routledge Handbooks in Applied Ethics (New York: Routledge, 2017), http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=nlebk&AN=1581310&site=ehost-live&scope=site.

[55] "Indiscriminate Discrimination: A Correspondence Test for Ethnic Homophily in the Chicago Labor Market," *Labour Economics* 19, no. 6 (December 1, 2012): 824–32, https://doi.org/10.1016/j.labeco.2012.08.004.

[56] United States and John Podesta, eds., *Big Data: Seizing Opportunities, Preserving Values* (Washington: White House, Executive Office of the President, 2014).

[57] Steve Lohr, "Facial Recognition Is Accurate, If You're a White Guy," *The New York Times*, February 9, 2018, sec. Technology, https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

[58] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2016), https://papers.ssrn.com/abstract=2477899.

[59] Doug Criss CNN, "Judge Rules California Baker Doesn't Have to Make Wedding Cake for Same-Sex Couple," CNN, accessed March 5, 2018, https://www.cnn.com/2018/02/08/us/wedding-cake-ruling-trnd/index.html.

[60] Jon Russell, "Alibaba Debuts 'Smile to Pay' Facial Recognition Payments at KFC in China," *TechCrunch* (blog), accessed January 30, 2018, http://social.techcrunch.com/2017/09/03/alibaba-debuts-smile-to-pay/.

[61] Batya Friedman, Peter H. Kahn, and Alan Borning, "Value Sensitive Design and Information Systems," in *The Handbook of Information and Computer Ethics*, ed. Kenneth Einar Himma and Herman T. Tavani (Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008), 69–101, https://doi.org/10.1002/9780470281819.ch4.

[62] Terrell Bynum, "Computer and Information Ethics," August 14, 2001, https://plato.stanford.edu/archives/win2014/entries/ethics-computer/.

[63] Jackie SnowMar 7, 2017, and 3:15 Pm, "Brainlike Computers Are a Black Box. Scientists Are Finally Peering Inside," Science | AAAS, March 7, 2017, http://www.sciencemag.org/news/2017/03/brainlike-computers-are-black-box-scientists-are-finally-peering-inside.

[64] Jackie Snow, "We Are Starting to Peer inside 'Black Box' AI Algorithms," MIT Technology Review, accessed March 5, 2018, https://www.technologyreview.com/s/609338/new-research-aims-to-solve-the-problem-of-ai-bias-in-black-box-algorithms/.

[65] Sarah Tan et al., "Auditing Black-Box Models Using Transparent Model Distillation With Side Information," *ArXiv:1710.06169 [Cs, Stat]*, October 17, 2017, http://arxiv.org/abs/1710.06169.

[66] Julia Angwin Jeff Larson, "How We Analyzed the COMPAS Recidivism Algorithm," text/html, ProPublica, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[67] Tom Simonite, "How Coders Are Fighting Bias in Facial Recognition Software," WIRED, March 29, 2018, https://www.wired.com/story/how-coders-are-fighting-bias-in-facial-recognition-software/.

[68] Laura Hudson, "Technology Is Biased Too. How Do We Fix It?," *FiveThirtyEight* (blog), July 20, 2017, https://fivethirtyeight.com/features/technology-is-biased-too-how-do-we-fix-it/.

[69] Hee Jung Ryu, Margaret Mitchell, and Hartwig Adam, "Improving Smiling Detection with Race and Gender Diversity," *ArXiv:1712.00193 [Cs]*, December 1, 2017, http://arxiv.org/abs/1712.00193.

[70] "45 CFR 46," Text, HHS.gov, February 16, 2016, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html.

[71] "What Is the Definition of Minimal Risk?," Research Office, May 25, 2012, http://research.oregonstate.edu/irb/frequently-asked-questions/what-definition-minimal-risk.

[72]李建华, and 冯昊青. "核伦理学研究的转型与走向." 哲学研究 4 (2008): 110–17.

[73] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, First edition (New York: Crown, 2016).

[74] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks," *Proceedings of the National Academy of Sciences* 111, no. 24 (June 17, 2014): 8788–90, https://doi.org/10.1073/pnas.1320040111.

[75] Danah Boyd, "Untangling Research and Practice: What Facebook's 'Emotional Contagion' Study Teaches Us," *Research Ethics* 12, no. 1 (January 1, 2016): 4–13, https://doi.org/10.1177/1747016115583379.

[76] Stanford Law Review, "Consumer Subject Review Boards," Stanford Law Review, September 3, 2013, https://www.stanfordlawreview.org/online/privacy-and-big-data-consumer-subject-review-boards/.

[77] Zachary M. Schrag, "The Case against Ethics Review in the Social Sciences," *Research Ethics* 7, no. 4 (December 1, 2011): 120–31, https://doi.org/10.1177/174701611100700402.

[78] Molly Jackman and Lauri Kanerva, "Evolving the IRB: Building Robust Review for Industry Research," *Washington and Lee Law Review Online* 72, no. 3 (June 14, 2016): 442.

[79] Kate Crawford and Ryan Calo, "There Is a Blind Spot in AI Research," *Nature News* 538, no. 7625 (October 20, 2016): 311, https://doi.org/10.1038/538311a.

[80] Nicholas Diakopoulos and Sorelle Friedler, "We Need to Hold Algorithms Accountable—here's How to Do It," MIT Technology Review, accessed April 22, 2018, https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/.

[81] "AI Principles," Future of Life Institute, accessed March 20, 2018, https://futureoflife.org/ai-principles/.

[82] IAF Big Data Ethics Initiative, "Unified Ethical Frame for Big Data Analysis," March

2015.