A Virtue Ethics Analysis of Racial Bias in Google Facial Recognition Technology

STS Research Paper
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Justin Ngo

February 28, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: _____Justin Ngo_____

Approved: _____ Date _____
Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Introduction

Facial recognition software has become more prevalent in everyday life. The technology is used to unlock smartphones, and on social media platforms to identify and tag people in photos. Google Photos is one technology that uses facial recognition software to recognize faces of humans and animals in pictures, or objects and landmarks, grouping similar ones together. The algorithm labels groups with descriptive names based on what it recognizes in the pictures. Despite its growing use in daily life, facial recognition software has been increasingly observed to perpetuate racism. In 2015, Google Photos was condemned because its auto-labeling software tagged two Black friends as "gorillas". The mistake reflected a racist depiction which dates back centuries to scientific racism and the association of Black people with simians in the Great Chain of Being. In recent years, researchers and engineers have recognized the presence of racial bias in machine learning and have examined its harmful effects on racial minorities. However, most research fails to question the moral valence of companies pushing facial recognition software into society without heavily considering its racist ramifications. If software that reinforces racial bias continues to be used without correcting the underlying issues in its algorithms, then racial minorities will continue to be disproportionately harmed.

I will investigate the racial bias encoded in Google Photos facial recognition software through the ethical framework of virtue ethics to explain that Google and its software engineers, who built its facial recognition technology, are morally responsible for perpetuating racism. I will do this by examining the cardinal virtue of justice and its relation to fairness, equity, freedom from bias, and the responsibility of engineers to uphold this virtue (van de Poel & Royakkers, 2011). Through the lens of virtue ethics, I will review Google's use of racially biased data and failure to fix the issue in its underlying algorithms.

Background

Google LLC is an American multinational technology company that specializes in Internet-related services and products. The company is a leader in the field of artificial intelligence and machine learning (Barr, 2015). Google Photos is a photo sharing and storage service developed by Google. It was announced in May 2015 and branched off from Google+, the company's former social network. The service uses machine learning to analyze photos, identifying various visual features and subjects within them. Users can search for anything in photos, with the service returning results from three major categories: people, places, and things. (Google Photos). The computer vision of Google Photos recognizes faces, grouping similar ones together; geographic landmarks; and subject matter (Simonite, 2018). Different forms of machine learning in the Google Photos service allow recognition of photo contents, automatic generation of albums, animation of similar photos into quick videos, and improvement in the quality of photos and videos (Google Photos).

Literature Review

In recent years, scholars have investigated the impact of racial bias in facial recognition software. In the United States, discriminatory technology is a lingering effect of the country’s history of racism. Racially biased data collected in the past is being used to build facial recognition algorithms, and these algorithms are appearing in an increasing number of new technologies. The following analyses discuss racial bias in these technologies, but not how it exhibits a lapse in the practice of virtue ethics on the part of engineers to fix racism in facial recognition software.

In her book, *Race After Technology*, author Ruha Benjamin examines how emerging technologies can reinforce white supremacy and deepen social inequality. She explains that algorithms have the potential to hide, speed up, and deepen discrimination while appearing facially neutral and even benevolent compared to past, explicit racism. Benjamin builds upon the argument of race scholars that ill intent is not always a feature of racism, and that technology can reflect and reproduce existing racial inequities without necessarily having creators who hold racist intentions. Furthermore, Benjamin points out that in the face of the discriminatory effects of technology, if the creators and those in power of the technology continue to use the technology, they are perpetuating a racist system (Benjamin, 2019).

In his paper "Understanding Bias in Facial Recognition Technologies," David Leslie describes how historical patterns of discrimination made their way into the design and implementation of facial recognition software. Leslie explains that the algorithms in facial recognition technology can be racially biased because of the convergence of two complex issues: the culturally entrenched legacies of historical racism and white male privilege in visual reproduction technologies, and the development of new sources of bias and discrimination arising from novel sociotechnical contexts of algorithmic design. He claims that algorithmic bias in facial recognition technology involves a carry-over of certain discriminatory assumptions from the history of photography into computer vision. Leslie explains that in the past, cameras were built to prioritize capturing white skin and did not account for the difference in contrast in darker skin. Leslie adds that human biases have crept into model design choices through massive datasets. These datasets tend to overrepresent dominant groups and marginalize people of color but continue to be fed into data-driven machine learning algorithms. Leslie includes research from 2002 to 2019 demonstrating significant racial and gender biases in widely used facial

recognition algorithms. There was a dramatic accuracy differential in applications of facial recognition software to distinguish between gender, age, and racial groups. The research revealed that historically marginalized and non-dominant subpopulations suffered from the highest levels of misidentification and the greatest performance drops (Leslie, 2020).

In their paper "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Joy Buolamwini and Timnit Gebru build upon research demonstrating that machine learning algorithms can discriminate based on classes like race and gender. In their paper, they present an approach to evaluate bias present in automated facial recognition algorithms and datasets with respect to phenotypic subgroups. Buolamwini and Gebru audited commercial gender classification systems for performance disparities. Their audit showed that the rate of misclassification for darker-skinned women was 35 times or more higher than for white men with error rates of up to 34:7%. They found that the datasets these commercial systems produced were overwhelmingly composed of lighter-skinned subjects. In addition, their work highlighted the compounding effects of intersectional discrimination in facial recognition technologies trained on largescale datasets. (Buolamwini & Gebru, 2018).

Technology often hides and deepens discrimination while appearing to be facially neutral or benevolent when compared to past, explicit racism. Historical patterns of discrimination have evidently made their way into the design and implementation of facial recognition software. As a result, racism is not only magnified, but also buried under layers of digital denial because technological advances are sold as morally superior and perceived to rise above human bias—even though the technology could not exist without data produced through a long history of exclusion and discrimination (Benjamin, 2019). Structural racism exists in our society and as a result, the data used to build the algorithms in facial recognition technology can reflect and

amplify the issue. I will deploy the framework of virtue ethics to question the morality of Google and its developers in relying on racially biased data to not only create, but continue using its facial recognition software without fixing the underlying issues in the technology.

Conceptual Framework

Google Photos facial recognition technology will first be addressed by examining the racial bias in its algorithms, followed by a discussion about what makes the technology and its creators racist, and why. The ethical framework of virtue ethics will be used to address the morality of Google and its software engineers responsible for building its facial recognition technology. Virtue ethics, developed by Aristotle, evaluates the character of the moral actor through their actions (van de Poel & Royakkers, 2011). In this case, the actor and actions being examined are Google software engineers, and their encoding and lack of greater action in addressing race-related flaws in the underlying algorithms used in Google Photos. The framework of virtue ethics focuses on the qualities of excellence or "virtue" that people should practice in acting morally and attaining the telos, or end goal, of the good life. Aristotle describes the good life, also known as eudaimonia, as being one lived in accord with nature— the way humans are "meant" to live. Humans are rational by nature, and as a result they should use reason to determine how to live morally or virtuously. However, Aristotle asserts that virtues are not innate, but can be learned from examples and through practice. In addition, it takes moral skill to discern which virtue is required in a certain situation. To be virtuous, moral skill must be put into action or performed when an opportunity to be moral arises, with the proper motivation, and goal in mind for the good. Some cardinal virtues of justice described by Aristotle, Cicero, and Plato include fairness, equity, and freedom from bias.

Justice can be defined as the equity that ensures that all people are on a level playing field and can be viewed as balance between selflessness and selfishness. Aristotle describes equity as "justice which lies beyond the written law" (Shiner, 1994). The virtue of equity is required for justice and is used in resolving disparities between written law and societal standards. In Computer Science, The ACM Code of Ethics and Professional Conduct outlines how computing professionals can act to uphold justice. The ACM Code outlines responsibilities of software engineers including, but not limited to, avoiding harm, being fair and acting not to discriminate, and giving comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks (ACM, 2021). Using biased data to create facial recognition technology does not align with the virtue of justice and freedom from bias because the technology fails to treat all members of a population equally and fails to treat marginalized members of society equitably. It can be concluded that individuals involved in the development of technology, who create and continue to use technology based on biased data, lack a complete understanding of justice, fairness, and equity in engineering practice

In addition, professional "cardinal" virtues include virtues essential to a profession that define the character of a moral professional in a particular field. In this case, the professionals are software engineers. Accordingly, I will use Pritchard's list of "Virtues for Morally Responsible Engineers" as a set of relevant virtues essential to engineering practice. These virtues include, but are not limited to, perseverance in addressing design issues, committing to objectivity, being fair, commitment to quality, and openness to correction Pritchard states that lacking any one of these virtues is sufficient to detract from responsible engineering practice (Pritchard, 2001).

Through the lens of virtue ethics, I will question whether Google Photos facial recognition algorithm can be deemed just and fair based on the values outlined by Aristotle, The

ACM Code of Ethics and Professional Conduct, and Pritchard's "Virtues for Morally Responsible Engineers". I will do this by investigating the impacts of the facial recognition technology to determine whether the technology is unbiased and its creator's commitment to quality and correction of mistakes. Through this, I will use virtue ethics to determine whether Google and its software engineers can be held morally responsible for perpetuating racism.

Analysis

In computer science, The ACM Code of Ethics and Professional Conduct outlines how computing professionals can act responsibly to uphold justice and practice virtue ethics. To act responsibly and consistently support the public good under The ACM Code, computing professionals should reflect upon the wider impacts of their work in avoiding harm, being fair and acting not to discriminate, and giving comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks (ACM, 2021). In addition, according to Pritchard's virtues for morally responsible engineers, virtue ethics in engineering practice addresses questions of character by establishing principles of perseverance in addressing design issues, committing to objectivity, being fair, commitment to quality, and openness to correction (Pritchard, 2001). The Google Photos software engineers exist at the intersection of computing and engineering. Thus, these engineers are morally responsible for upholding The ACM Code and engineering virtue ethics. These software engineers are responsible for perpetuating a racist system in commercializing discriminatory technology created by racially biased data, and for failing to persevere in correcting the underlying design issues maintaining such racist effects. As explained by Benjamin, technology can reflect and reproduce existing racial inequities without necessarily having creators who hold racist intentions. In addition, in the

face of the discriminatory effects of technology, Benjamin points out that if the creators and those in power of the technology continue to use the technology, they are perpetuating a racist system (Benjamin, 2019). Furthermore, historical patterns of discrimination have made their way into the design and implementation of facial recognition software (Leslie, 2020). As a result, there is significant bias in datasets used in the creation of facial recognition algorithms. Therefore, the software built from that data cannot be considered just knowing that the data available and used to create it contains bias. Facial recognition has a significant effect and use in everyday society and is capable of amplifying racism. In my analysis, I will demonstrate the misconduct of Google software engineers in failing to fix the company's racist facial recognition technology as a direct result of a lack of the cardinal virtue of justice and practice of engineering virtue ethics. The examples I use to demonstrate failures in these areas are Google's use of racially biased data to build its facial recognition algorithms, and Google's software engineers failing to correct the underlying issue in the algorithms years after its initial discovery.

Using Racially Biased Data

Google software engineers are responsible for perpetuating a racist system because the company commercialized discriminatory facial recognition technology built using racially biased data. In 2015, when Jacky Alcine, a resident of Brooklyn, New York logged onto Google Photos, he was shocked to find that the technology had created an album titled "Gorillas," in which the facial recognition software categorized him and his friend as primates. Immediately, Alcine posted on Twitter about the offensive incident— prompting over 1,000 re-tweets and an online discussion on the shocking situation. As explained by Benjamin, technology can reflect and reproduce existing racial inequities without necessarily having creators who hold racist

intentions (Benjamin, 2019). According to Leslie, the software built from that data cannot be considered just knowing that the data available and used to create it contains bias. Furthermore, as noted by Barr, getting facial recognition technology right is increasingly important as it is used for more everyday tasks. For instance, Google's self-driving cars, which are being tested on public roads, use the same facial recognition technology to recognize objects and decide whether to stop, avoid, or continue. Photo recognition algorithms need to be more accurate and require an understanding cultural sensitivity, that are important to humans, before being integrated into society (Barr, 2015).

Google stated in the past that as more images are loaded into Google Photos and more people correct mistaken tags, its algorithms will get better at categorizing photos (Barr, 2015). However, this is not an acceptable explanation to the problematic algorithm which continues to be used by the company. The incident was a clear demonstration of racially biased data influencing the algorithm at the core of Google Photos facial recognition software. In addition, it revealed a lack of consideration for people with darker skin before release of the technology into society. According to The ACM Code of Ethics and Professional Conduct, computing professionals are responsible for avoiding harm, and being fair and acting not to discriminate. The racial bias in the datasets that Google used and in its facial recognition software do not align with the virtue of justice and freedom from bias because it fails to treat Black people equally and equitably. Furthermore, Google commercialized the product, and the technology benefited its software engineers at the expense of reflecting racism. It can be concluded that the software engineers involved in the development of the Google Photos facial recognition technology, who create and continue to use software based on racially biased data, lack a complete understanding of justice, fairness, and equity in engineering practice.

Failure to Fix the Underlying Algorithm

Google software engineers are responsible for perpetuating racism because the company to has failed to persevere in correcting the underlying design issues in its facial recognition software. After the incident in 2015, in which the Google Photos algorithm tagged a group of Black friends as "Gorillas," the company simply removed the "Gorilla" category from being a possible category, so that the specific suggestion would no longer appear (Barr, 2015). However, this solution does not address the root of the problem because the algorithm continues to demonstrate racial bias and groups people with darker skin in the same way— even if not necessarily as "Gorillas". In addition, now the algorithm will not correctly identify gorillas in pictures where the animal is present (Simonite, 2018).

Six years later, despite Google promising a fix, it has not announced one for the underlying algorithm. Google and its software engineers have failed in demonstrating perseverance in solving this issue because they have not continued to try and fix the algorithm before letting it be used. In addition, keeping the technology up for commercial use without fixing the underlying issue behind the racist categorization is not an acceptable solution. In a test conducted by WIRED in 2018 attempting to assess Google Photos' view of people, WIRED also uploaded a collection of more than 10,000 images used in facial-recognition research. The search term "African American" turned up only an image of grazing antelope. Typing "black man," "black woman," or "black person," caused Google's system to return black-and-white images of people, correctly sorted by gender, but not filtered by race. The only search terms with results that appeared to select for people with darker skin tones were "afro" and "African," although results were mixed (Simonite, 2018). According to Pritchard's "Virtues for Morally Responsible

Engineers," engineers are responsible for being open to correcting and persevering to fix unfair designs (Pritchard, 2001). The Google Photos software engineers, in failing to address the root of the issue, have not demonstrated a commitment to quality and correction of mistakes in line with that of a morally responsible engineer. Furthermore, in allowing the technology to remain open for commercial use without fixing the underlying issue in its facial recognition software for over half a decade, Google and its software engineers are morally responsible for perpetuating racism.

Conclusion

Although the specific code and datasets used in Google Photos facial recognition software are not available for public view, it is possible to make informed judgments about the character of the software engineers based on the decisions made during development of this software. Through the lens of virtue ethics, I have argued that the decisions made by the software engineers working on Google Photos reveal significant failures with respect to virtues necessary for morally responsible engineers: using racially biased data to build technology for commercial use and failing to persevere in correcting the underlying racist effects of its algorithm. Using a virtue ethics framework, the actions of Google software engineers are deemed immoral as they fail to demonstrate the characteristics that a virtuous computing professional and engineer would in the same circumstances. Examples from academic literature on racial bias in facial recognition technology prove that the virtue of justice cannot exist with algorithms built on racially biased data and the failure to fix underlying issues before allowing commercial use.

Software engineers are not just responsible for writing code and technical designs; they are often responsible for being just in their designs. Consideration of virtue ethics is critical to

engineering practice. Judging decisions with respect to these virtues provides a map to navigate complex situations and ultimately build a better world.

Word Count: 3287

References

ACM (2021, March 11). *ACM Code of Ethics and Professional Conduct* https://www.acm.org/code-of-ethics

Barr, A. (2015, July 1). G*oogle mistakenly tags Black people as 'Gorillas,' showing limits of algorithms*. The Wall Street Journal. https://www.wsj.com/articles/BL-DGB-42522

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

Buolamwini, J., & Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Google Photos. Retrieved March 11, 2022, from https://photos.google.com/

Leslie, D. (2020). *Understanding Bias in Facial Recognition Technologies: An Explainer*. The Alan Turing Institute. https://doi.org/10.5281/zenodo.4050457

Pritchard, M. (2001). Responsible engineering: The importance of character and imagination. *Science and Engineering Ethics*, 7(3), 391–402.

Simonite, T. (2018, January 11). *When it comes to gorillas, Google photos remains blind*. Wired. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Shiner, R. (1994). *Aristotle's Theory of Equity*. Loyola Law School. https://pdfs.semanticscholar.org/4ff2/32abf89f22c861057acbf9ea64a7016630a8.pdf

van de Poel, I., & Royakkers, L. (2011). *Ethics, Technology, and Engineering: An Introduction*. Blackwell Publishing.