

5-2022

Bias in Artificial Intelligence: The Morality and Motivation Behind the Algorithm

Avery Freeman
William & Mary

Follow this and additional works at: <https://scholarworks.wm.edu/honorstheses>



Part of the [Computer and Systems Architecture Commons](#), [Psychology Commons](#), and the [Race and Ethnicity Commons](#)

Recommended Citation

Freeman, Avery, "Bias in Artificial Intelligence: The Morality and Motivation Behind the Algorithm" (2022). *Undergraduate Honors Theses*. William & Mary. Paper 1807.
<https://scholarworks.wm.edu/honorstheses/1807>

This Honors Thesis -- Open Access is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

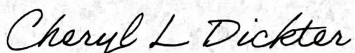
Bias in Artificial Intelligence: The Morality and Motivation Behind the Algorithm

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Science in Psychological Sciences from
William & Mary

by

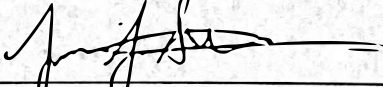
Avery M. Freeman

Accepted for Honors
(Honors, High Honors, Highest Honors)


Cheryl Dickter


Ashleigh Everhardt Queen


Reya Farber


Jessica Stephens


Paul Davies

Williamsburg, VA
May 12, 2022

RUNNING HEAD: BIAS IN ARTIFICIAL INTELLIGENCE

Bias in Artificial Intelligence: The Morality and Motivation Behind the Algorithm

Avery M. Freeman

College of William and Mary

Abstract

More than 180 cognitive biases have been identified in humans, and these biases relate to feelings towards a person or a group based on perceived group membership (Dilmegani, 2020). The development of artificial intelligence has fallen into the hands of engineers and statisticians, people who work within fields that have well-established race and gender diversity disparities (Panch et al., 2019). Thus, it is no surprise that the aforementioned biases have made their way into the algorithms behind artificial intelligence. The current study explored how participants' pre-existing biases and level of outgroup contact have the potential to affect their decision-making pertaining to the development of artificial intelligence algorithms. College student participants viewed pictures of faces on a computer screen varying in their racial identity (i.e., Black and White) and were asked to make decisions relevant to situations that artificial intelligence algorithms are being programmed to do. Eye tracking was recorded to investigate implicit attention to the faces. Results indicated that eye tracking patterns differed as a function of race when people were making decisions about hitting pedestrians while driving and during facial recognition. Outgroup contact did not moderate these effects. This study has implications for how implicit patterns of attention may present in human decision-making in the context of programming artificial intelligence.

Introduction

Previous research has demonstrated that racial bias is manifested in the algorithms used to produce artificial intelligence (AI) powered technology. For instance, with the commercial algorithms widely used by the United States health care system for the purpose of guiding health care decisions, more than 200 million United States citizens have experienced said bias (Dilmegani, 2022). Although these algorithms are designed for the purpose of predicting which patients would most likely need medical care and used previous patients' healthcare spending as a proxy for medical needs, because the algorithm relies on a faulty metric for determining need (Dilmegani, 2022), Black patients who were sicker were assigned the same level of risk as White patients who were not as sick (Obermeyer, Powers, & Mullainathan, 2019). This is an example of a poor interpretation of historical data due to the fact that income and race are highly correlated metrics. Therefore, making assumptions based on only one variable of correlated metrics compelled the algorithm to produce inaccurate results. Reformulating this algorithm would increase the percentage of Black patients receiving additional help from 17.7 to 46.5% (Obermeyer, Powers, & Mullainathan, 2019).

Despite the fact that errors like this can have extensive negative impacts, such as inadequate access to health care, on already vulnerable populations, not much has been done to remedy these situations. "Imagine an algorithm that selects nursing candidates for a multi-specialty practice—but only selects White females. Consider a revolutionary test for skin cancer that does not work on [Black people] ... These are [additional examples of how] ungoverned artificial intelligence might perpetuate bias (Nelson, 2019, p. 220)." In order to understand what drives this bias, it is critical to understand how these algorithms get programmed and who is behind them. The primary way in which biased algorithms come to fruition is that they are given

data that reflects what is going on in society, not just at that moment, but also in the past. So, the inequalities that are and have been present in society are emulated by the algorithms (LibertiesEU, 2021). Broadly, an algorithm is a set of instructions that tells a computer how to interpret certain pieces of information, and from that information, make a decision. This process revolves around three primary steps: process of input, transformation, and output. Due to the non-transparent nature of this process, it is incredibly easy to apply unethical criteria without consumer knowledge (LibertiesEU, 2021).

Common misconceptions about the role that humans play in the development of algorithms has enabled the fact that human bias often acts as catalyst of algorithmic bias to fly under the radar (LibertiesEU, 2021). The algorithms themselves are not in control; there are people behind the scenes that create them and adjust them. Outputs pertaining to an algorithm can be biased by humans at various levels, from what data gets selected to how the programmer decides to build the algorithm. Thus, human bias has proven to be one of the hardest parts to remove in algorithmic bias (LibertiesEU, 2021). In order to address this detrimental impact of human biases on algorithms, it is critical to let AI work be guided by the tenets of transparency, trust, fairness, and privacy (Nelson, 2019).

Transparency requires AI authors to explain both what is behind an algorithm and its results, and what decisions they made and why. Trust begins with this transparency, in addition to verification and accountability. As a social construct, fairness requires social responsibility, and in the case of AI, this would mean ensuring that algorithms do not discriminate against people based on protected traits (age, gender, race, etc.). Lastly, privacy has to do with the relationship between AI and its users, and thus, the individual privacy of the user must be protected at all times (Nelson, 2019). Taking these tenets and putting them into action has the

potential to not only mitigate the effects of human bias in technology, but also to spread awareness about the issue.

AI bias is defined as an anomaly in the output of machine learning algorithms as a result of assumptions made during the development of the algorithm or prejudices in data that were used to train the algorithm. These biases occur due to two primary reasons: cognitive biases and lack of complete data (Dilmegani, 2022). More specifically, the bias is a reflection of the data that was chosen, the data blending methods, model construction practices, and how the results are applied and interpreted (Nelson, 2019). The term cognitive bias refers to unconscious errors in thinking that affect an individual's judgements and decisions, due to the brain attempting to simplify processing information about the world (Dilmegani, 2022). In the United States, implicit biases tend to present themselves as racial prejudice, stereotypes, and attitudes (Greenwald & Krieger, 2006).

Through various research endeavors, it has been proposed that “ethnocentrism and prejudice have their origins in the process of social categorization, when people subjectively classify others as members of their own group (in-group) or as members of another group (out-group) (Dovidio, Gurtman, Perdue, & Tyler, 1990, pg. 475).” Additionally, social identity theory suggests that an individual's need for positive self-esteem will guide perceivers to favor their in-groups in most comparisons with out-groups (Dovidio, Gurtman, Perdue, & Tyler, 1990). An example of stereotype activation is the fact that names and labels applied to people can subtly alter impressions of said people. Collective pronouns (we, us, ours, they, them, theirs) can be particularly powerful influences in both social categorization and perception. Research conducted by Dovidio, Gurtman, & Tyler (1990) showed that in-group and out-group

designating pronouns possess different evaluative valences. Therefore, they may produce automatic responses to positive and negative information (Dovidio et al., 1990).

Unlike explicit attitudes, which are demonstrated by the attitudes measured by standard self-report measures, implicit attitudes are evaluations that are automatically activated by the presence of the attitude object (Dovidio, Gaertner, & Kawakami, 2002). Most commonly, these research endeavors center themselves around the study of implicit biases redirected towards members of stigmatized groups (e.g., people of color, women, the LGBTQ+ community). Implicit and explicit attitudes have the potential to be consistent or inconsistent and they typically diverge for socially sensitive issues (Dovidio, Gaertner, & Kawakami, 2002). For example, a man who explicitly believes men and women are equally suited for careers outside the home could still implicitly distrust feedback from female co-workers, as well as hire equally qualified men over women. It would be reasonable to conclude that the aforementioned individual is demonstrating an implicit gender bias (Stanford Encyclopedia of Philosophy, 2019).

Measuring implicit bias goes beyond asking someone what they think about something by relying on explicit forms of questioning (Stanford Encyclopedia of Philosophy, 2019). In contrast to explicit bias, implicit attitudes influence responses that are more difficult to monitor and control (e.g., some non-verbal behaviors) or responses that people do not see as an indication of their attitude and thus do not attempt to control (Dovidio, Gaertner, & Kawakami, 2002). It is also important to note that implicit biases tend to be thought of as unconscious for several reasons: there is no phenomenology associated with the relevant mental states or dispositions, the agent is unaware of the content of the representations underlying their performance, the agent is unaware of the source of their implicit biases, the agent is unaware of the relations between their relevant states, and the agent might have different modes of awareness of their own mind

(Stanford Encyclopedia of Philosophy, 2019). Thus, pinpointing exactly where one's implicit biases are coming from can be quite a difficult task and are often measured through reaction time-based tasks or physiological measures such as eye tracking. Overall, both implicit bias and explicit bias predict different types of behaviors, which is why it is important to study both.

One factor that may affect an individual's bias, specifically racial bias, is intergroup contact. Both diversity in childhood and current contact has the ability to shape implicit racial bias across perceivers' racial group (Kubota, Peiso, Marcum, & Cloutier, 2017). Kubota et al. (2017) conducted two investigations where the participants completed an Implicit Association Test and a self-report measure of the racial diversity of their current and childhood contact. A reduced implicit pro-White racial bias was demonstrated when increased contact with Black individuals compared with White individuals occurred in childhood (Study 2) and currently (Studies 1 and 2). Additionally, for Black participants (Study 2), increased contact with Black individuals compared with White individuals resulted in reduced pro-White racial bias. Thus, it is plausible to conclude that diversity in contact over the course of one's life may have an impact on the expression of implicit racial bias. Furthermore, this relationship can be generalized across racial groups. Given the wealth of evidence suggesting that humans hold implicit biases that shape their judgments and decision-making and that humans are responsible for programming the algorithms that guide AI, it is important to examine the biases that affect judgments that contribute to the development of the algorithms.

The Current Study

The purpose of this study was to evaluate the decision-making process of participants as they were placed in the position of making similar implicit decisions that would go into programming an AI algorithm. In addition, the study aimed to examine whether their implicit

decision-making process would be affected based on their level of outgroup contact (during childhood and currently). To investigate this research question, eye tracking and mouse tracking were used, in addition to a survey, in order to assess outgroup contact as a moderator of implicit bias expression. The stimuli were presented on a computer screen, and they represented three different identities: race, gender, and age.

Although race was the demographic of interest, gender and age were used to explore potential intersectionality effects. However, no specific hypotheses related to gender and age were generated. Due to the fact that cognitive biases are quick and automatic, eye tracking can be used as a window into an individual's decision-making strategy. An area of interest at the intersection of eye tracking and the expression of implicit biases during decision-making is top-down visual attention (Matzen, Hass, & McNamara, 2014). Previous research conducted by Matzen, Hass, & McNamara (2014) has utilized eye tracking to breakdown the various components of top-down visual attention processing.

First, top-down visual attention is a reflection of an individual's goals. Therefore, an eye tracking task has the ability to determine how the participant is assessing the stimuli and how that assessment is related to their goals. Additionally, top-down attention is affected by cognitive load, working memory, past experiences, and past knowledge (Matzen, Hass, & McNamara, 2014). As a result, cognitive biases can be reflected in the participant's eye movements during a task. Lastly, top-down information creates a pre-attentive ranking of items to establish attentional priority (Matzen, Hass, & McNamara, 2014). In the current study, assessing how long participants spent looking at White faces compared to Black faces could provide insight as to how their attentional prioritization impact their decision-making process related to choosing a face based on the given scenario.

When it comes to making decisions, they are often complex and must be made with relatively little or ambiguous information. As a result, one seeks to resolve their decision conflict between multiple possible alternatives (Stillman, Shen, & Ferguson, 2018). This process can involve anything from evaluation to categorization, and thus, common themes across different domains may arise (social categorization, self-control, prejudice, etc.). Mouse-tracking is a valuable asset when it comes to exploring an individual's real-time decision conflict resolution. These movements made by participants are measured and rich and accessible data gets produced. Primarily, mouse tracking allows researchers to more accurately evaluate the magnitude of conflict present during a particular decision (Stillman, Shen, & Ferguson, 2018).

Decisions where categorization judgements can be influenced by stereotypes and prejudices are critical to explore. Not only can the process of categorization become clouded due to the fact that information in the world can be ambiguous, but it can also become clouded as a result of stereotype knowledge biasing categorization. Stoller and Freeman (2016) used a mouse tracking facial categorization task to explore how multiple social categories activate and resolve over hundreds of milliseconds during real-time categorization. During the task, the participant's hand movement trajectory was recorded en route to the selected response. For instance, the presentation of a Black female face tends to elicit an eye trajectory that initially deviates toward the male response because shared stereotypes exist between Black and male categories. Therefore, the perception of Black faces is biased towards male categorization. The overall findings of this study suggest that social-conceptual knowledge can systematically alter the representational structure of social categories at multiple levels of cortical processing. Thus, bias is reflected in visual perceptions (Stoller & Freeman, 2016).

There are many ways by which mouse tracking data can be analyzed. The most common approach involves quantifying the relative conflict present on a specific trial (Stillman, Shen, & Ferguson, 2018). Another option is to look at entropy and uncertainty, which means investigating the relative unpredictability that a given trajectory demonstrates (Stillman, Shen, & Ferguson, 2018). It can also be beneficial to explore how X-location, velocity, and acceleration profiles unfold during the task. More pointedly, drawing on research in dynamical systems, researchers can use these profiles to adjudicate between predictions of sequential versus dynamical systems, as well as to inspect the relative presence or lack of conflict (Stillman, Shen, & Ferguson, 2018). Lastly, integration times can be used to examine the temporal dynamics involved in predicting when in the trajectory the angle of movement is significantly influenced by the attributes of the stimuli (Stillman, Shen, & Ferguson, 2018).

The main hypothesis of the current study was that participants would show racial differences in how they implicitly process two faces of people with different racial identities (i.e., Black and White) presented on a computer screen and that these differences would be based on racial stereotypes in our society conveying positive traits about White people and negative traits about Black people (Devine, 1989). More specifically, they would spend more time looking at Black faces in the self-driving car scenario, White faces in the hiring scenario, and White faces in the facial recognition scenario. In addition, we expected that participants who had more outgroup contact over the course of their lifespan thus far would demonstrate less bias than those who have not experienced a significant amount of outgroup contact over the course of their lifespan thus far.

Method

Participants

36 participants were recruited from a medium-sized public liberal arts university in the Southeast and participated for one course credit. All procedures were approved by the university's Protection of Human Subjects Committee; informed consent was obtained from each participant.

Experimental Paradigm

Images of Black and White male and female faces were used as stimuli. The faces (Minear & Park, 2004) all had neutral expressions, were presented in black and white, and two age ranges were represented (younger and middle-aged). Younger was defined as 18-34 years old and middle-aged was defined as 35-49 years old. All of the faces were presented in pairs, with one face positioned on the left side of the screen and one face positioned on the right side of the screen; one of three variables was the differentiating factor between them: race, gender, or age. After being presented with a brief eye tracker calibration task and the overall instructions for the task, the participants were given three different scenarios. After each scenario, the participants saw a pair of faces and had to select one based on the scenario. Each task and each trial were presented to the participant in a random order. 22 total face pairs differing in race ($n=10$), gender ($n=6$), and age ($n=6$) were presented with each scenario for a total of 66 trials.

The three scenarios that the participants were presented with were related to the development of algorithms for a self-driving car, a hiring program, and a facial recognition software. The self-driving car scenario was selected due to the fact self-driving cars need to make quick decisions that could be influenced by race, akin to the challenges presented by the classic Trolley Problem. The Trolley Problem is a thought experiment where a person is presented with two situations with similar choices and potential consequences. The switch situation involves a runaway trolley driving down a track. This trolley will run into and kill five

workmen unless the observer flips the switch and directs the train down a track that will only kill one workman. In the bridge situation, the observer must decide whether or not to push a plump individual off a bridge to stop the train and save the five workmen (Roff, 2018). Although the Trolley Problem detracts from understanding how autonomous cars work and the control that humans have over their decision-making, it is a great foundation to build off of in terms of diving into the ethical components of artificial intelligence development (Roff, 2018).

Amazon's biased recruiting tool was the inspiration behind the hiring program scenario. With the goal of automating the recruiting process in mind, Amazon began an artificial intelligence project in 2014 (Dilmegani, 2022). The project was exclusively based on reviewing job applicants' resumes and rating them using AI powered algorithms. Thus, recruiters would not have to spend an excessive amount of time doing manual resume screening tasks. Unfortunately, by 2015, Amazon uncovered that their system was not ranking the candidates fairly and it showed bias against women (Dilmegani, 2022). Bias expression such as this was made possible because Amazon used data from the last ten years to train their model. Male dominance within the tech industry meant that 60% of Amazon's employees at the time this data was collected were men (Dilmegani, 2022). The system then inaccurately learned that men were the preferential candidates for hiring. Historical data riddled with bias against women, in addition to the biases of those who selected this data set, were made a part of the algorithm and penalized resumes that included the word "women's" (ex. women's soccer captain). As a result, Amazon ceased their use of the algorithm (Dilmegani, 2022).

While facial recognition technology is beneficial when it comes to helping us unlock our phones or tagging us in pictures on social media, it is also used for more serious purposes by law enforcement, for airport passenger screening, and when making employment and housing

decisions. So, it was important to include a scenario related to this technology as a part of this study (Najibi, 2020). Despite that fact that facial recognition algorithms claim to be over 90% accurate, this accuracy rate is not universal. For instance, more and more research is revealing that there are divergent error rates across demographic groups (Najibi, 2020). The Gender Shades project shed light on discrepancies in the classification accuracy of facial recognition technologies for different skin tones and sexes. Overall, the algorithms consistently provided the poorest accuracy for darker-skinned females and the highest for lighter-skinned males (Najibi, 2020).

Racial discrimination related to facial recognition and law enforcement dates back to the 18th century, when New York put “lantern laws” into place that required enslaved people to carry lanterns at night so that they were publicly visible (Najibi, 2020). Today, facial recognition can be used by law enforcement to target marginalized populations, such as people of color or undocumented immigrants. Broadly, facial recognition software has the potential to compromise one’s privacy, freedom of expression, and freedom of association and due process. Yet, for Black Americans this is not a new threat. The FBI has long history of surveilling prominent black activists, and law enforcement in general has a long history of using surveillance technology to carry out targeted abuse towards the Black community (Najibi, 2020). Facial recognition algorithms that are inherently biased can misidentify subjects, which would help authorities to continue to incarcerate innocent Black Americans (Najibi, 2020).

Each scenario was written in a way that would not guide the participant to choose a certain race, gender, or age over another. They had an unlimited amount of time to read the instruction screen, to read the scenarios, and to choose a face to select. An example of one of the scenarios that was presented during the task is “You are responsible for developing the algorithm

of a self-driving car. In the instance of an accident, the car must decide which of two pedestrians must be struck. As the developer, you will be presented with several faces and tasked with deciding which pedestrian would be struck in the accident. Your input during this task will directly affect the programming of the vehicles.” Once the participant completed the task, they were asked to complete a survey.

Questionnaires

Close Friendships Questionnaire. The Close Friendships Questionnaire is used to assess the participant’s familiarity with the outgroup. Adapted from Greenwald, McGee, and Schwartz (1998), the first question asks the participant to list the initials of their 20 closest friends. They are then shown a list of those initials and asked to respond to two follow-up questions. The first follow-up questions ask how many of their friends on the list are White, and the second follow-up questions asks how many of their friends on the list are Black.

Feelings Thermometer. The feelings thermometer (Haddock, Zanna, & Esses, 1993) was presented to participants to evaluate participants’ overall attitude towards Blacks and Whites. If the participants had a favorable attitude toward the group, they were told to give that group a score between 50 degrees and 100 degrees, depending on how favorable they are to toward them. If the participant’s attitude was unfavorable, they were asked to provide a score between 0 and 50 degrees, depending on how unfavorable they are toward them. However, the participant was also informed that they are not restricted to the numbers indicated (60, 70, 80, etc.) and could use any number between 0 and 100 degrees.

Social Contact. The Social Contact questionnaire (Walker et al., 2008) asked participants to indicate how much they agreed with three different statements using “strongly disagree,” “sort of disagree,” “not sure,” “sort of agree,” and “strongly agree.” For example, “I often hang out with

Black people.” Participants indicated their agreement for the statements for Black and White individuals. The purpose of these questions was to establish the degree of social contact the participant has had with ingroup members compared to outgroup members.

Individuating Experience. The Individuating Experience questionnaire (Walker et al., 2008) asked the participant to indicate how often they participated in several activities using “never,” “rarely,” “once in a while,” “sometimes,” and “frequently.” For example, “how often do you spend time with Black friends at their place?” They were asked to answer this two times: once where the activities involved ingroup members (White) and once where the activities involved outgroup members (Black).

Cross-Group Friendship Measure. The Cross-Group Friendship Measure looked at how many friends the participant has in college that are Black (outgroup) compared to White (ingroup), and how often they spend time with their Black friends in college compared to their White friends. To indicate the number of friends, participants could select “none,” “one,” “two to four,” “five to ten,” or “over ten.” To indicate how much time they spend with each group of friends, the participants could select “never,” “occasionally,” “sometimes,” “often,” or “very often.”

Positive and Negative Contact (Direct and Indirect). These positive and negative contact questions not only assessed the experiences of the participant themselves (Wolfer et al., 2017), but they also assessed the experiences of the participant’s friends (Mazziotta et al., 2015). First, participants rated the frequency of their positive and negative experiences with Whites and Blacks using “never,” “occasionally,” “sometimes,” or “often.” Next, they used the same options to rate the frequency of their friend’s positive and negative experiences with Whites and Blacks. For example, “how often are your experiences with Blacks positive?”

Current Everyday Experiences with Members of a Minority Group (Whites and Blacks).

These questionnaires asked participants to reflect on their current everyday experiences with both ingroup and outgroup members. To answer each of the questions, participants were able to enter a number. For example, “how many people have you dated that are Black? How many people have you dated that are White?”

Identity with Ingroup and Outgroup (Cao et al.). The Identity with Ingroup and Outgroup questionnaire had participants provide ratings on a 5-point scale. On this scale, 1 was not similar at all and 5 was very similar.

Extended Contact (Capozza et al, 2014; Turner, Hewstone, Voci, & Vonofakou, 2008). The extended contact measure utilized a 5-point scale to assess the participant’s level of extended contact with outgroup members (Blacks). A 1 indicated “none”, a 2 indicated “a few”, a 3 indicated “about half,” a 4 indicated “more than half, and 5 indicated “most.” For example, “how many White people do you know that have friends who are Black?”

Perceived Threat (Tausch et al., 2007). In order to respond to the questions, participants read the following: “if you were the only member of your group and you were interacting with people from the other [outgroup] group, e.g., talking to them, working on a project with them, how would you feel?” Then, participants rated the extent to which they would feel nervous, anxious, comfortable, awkward, safe, and at ease using a 7-point scale. This scale ranged from 1 being “not at all” to 7 being “extremely.”

Past Experiences with Members of a Minority Group (Black). The Past Experiences with Members of a Minority Group questionnaire explored the participants’ past experiences with Blacks. The participants were able to enter numbers to indicate an amount or percentage to

answer each question. For example, “estimate the approximate percentage of Black students in your high school.”

Inclusion of the Outgroup in Self (Turner et al., 2008). Using a pictorial item (based on Aron et al.’s, 19922), inclusion of the outgroup in the self was measured. This item was comprised of seven pairs of overlapping circles, for which participants were asked to indicate the nature of their relationship with the outgroup. The greater the overlap between the circles, the greater the inclusion of the outgroup in self. The same was done to measure the inclusion of the ingroup in the self.

Relationship with Outgroup. The Relationship with the Outgroup questionnaire (Capozza et al, 2014) asked participants to use a 7-point scale to rate the following statement: “my identity, in a sense, also includes Black’s identity. A 1 was used to mean “not at all” and a 7 was used to mean “definitely true.”

Results

The final sample consisted of 36 individuals ranging in age from 18 to 21. Participants self-reported their race as follows: 1 Black, 4 Asian, 2 Hispanic, 3 Biracial, and 26 White. 25 of the participants self-identified as female, 10 participants self-identified as male, and 1 participant self-identified as non-binary. For this thesis, only eye tracking data (i.e., not mouse tracking) analyses will be described.

Eye Tracking Data Processing

Due to the fact that each participant got a break halfway through each scenario, there was a part I and a part II for each scenario when the data file was created; thus, the first and second parts of each of the scenarios were combined. Once this was completed, within each scenario, all of the conditions where race was the differentiating factor were isolated. For each condition, the

left image was area of interest number one, and the right image was area of interest number two. What faces were presented, and which image was on the right, and which was on the left was indicated by the condition label. In total, each scenario had 8 conditions that differed based on race.

Using these 8 conditions, proportions were then generated for each participant related to each scenario and how much time they spent looking at the Black vs White faces. To do so, the sum of all conditions where the Black faces were the area of interest (2 conditions on the right and 2 conditions on the left) was divided by the sum of all of the conditions total (8). The same was then done for all of the conditions where the White faces were the area of interest (2 conditions on the right and 2 conditions on the left). At the end of this process, each participant had a total of 6 percentages, 2 for each scenario, one being for the amount of time they spent looking at the Black faces and one being for the amount of time they spent looking at the White faces.

Data Analysis Strategy

In order to examine whether eye tracking patterns differed as a function of race we conducted a repeated measure analysis of variance (ANOVA) with race as the within-subjects variable for each of the scenarios. To examine whether eye tracking patterns to the different faces were moderated by individual difference variables, mixed model ANOVAs were conducted for each of the three scenarios.

Driving Task

A repeated measures ANOVA yielded a significant effect of race, $F(1,35) = 67.06$, $p < .001$, $\eta_p^2 = .657$, such that participants had more fixations to the White target ($M = .56$, $SE = .01$) than the Black target ($M = .44$, $SE = .01$). The mixed model ANOVA did not yield a

significant race by social contact interaction, $F(1,24) = 4.50, p = .674, \eta_p^2 = .158$. The mixed model ANOVA did not yield a significant race by college Black contact interaction, $F(1,23) = 6.135, p = .416, \eta_p^2 = .211$. The mixed model ANOVA did not yield a significant race by positive Black contact, $F(1,24) = 4.79, p = .187, \eta_p^2 = .166$. The mixed model ANOVA did not yield a significant race by negative Black contact interaction, $F(1,24) = 3.03, p = .608, \eta_p^2 = .112$.

Hiring Task

A repeated measures ANOVA did not yield a significant effect of race, $F(1,35) = 2.06, p = .160, \eta_p^2 = 0.56$. The mixed model ANOVA did not yield a significant race by social contact interaction, $F(1,24) = .05, p = .914, \eta_p^2 = .002$. The mixed model ANOVA did not yield a significant race by college Black contact interaction, $F(1,23) = .46, p = .284, \eta_p^2 = .020$. The mixed model ANOVA did not yield a significant race by positive Black contact, $F(1,24) = 1.86, p = .239, \eta_p^2 = .072$. The mixed model ANOVA did not yield a significant race by negative Black contact interaction, $F(1,24) = 2.54, p = .040, \eta_p^2 = .096$.

Facial Recognition Task

A repeated measures ANOVA yielded a significant effect of race, $F(1,35) = 63.75, p < .001, \eta_p^2 = .646$, such that participants had more fixations to the White target ($M = .62, SE = .01$) than Black target ($M = .47, SE = .01$). The mixed model ANOVA did not yield a significant race by social contact interaction, $F(1,24) = 5.27, p = .308, \eta_p^2 = .180$. The mixed model ANOVA did not yield a significant race by college Black contact interaction, $F(1,23) = 7.35, p = .189, \eta_p^2 = .242$. The mixed model ANOVA did not yield a significant race by positive Black contact, $F(1,24) = .52, p = .953, \eta_p^2 = .021$. The mixed model ANOVA did not yield a significant race by negative Black contact interaction, $F(1,24) = 1.28, p = .456, \eta_p^2 = .050$.

Discussion

This study examined the effect of implicit bias as measured by eye tracking patterns during a face selection task in a primarily White sample. Results indicated that there was a significant effect of race related to both the driving and the facial recognition scenarios such that participants attended more to the White faces than the Black faces. This study has implications for how implicit patterns of attention may present in human decision-making in the context of programming AI (Silberg & Manyika, 2019). Understanding decision-making differences with faces of different races, genders, and ages with a majority White sample will help inform how those who are most commonly in the position of influencing AI express biases and, subsequently, what can be done to eradicate said biases.

Our hypothesis that participants would show racial differences in how they implicitly process two faces of people with different racial identities (i.e., Black and White) presented on a computer screen and that these differences would be based on racial stereotypes in our society conveying positive traits about White people and negative traits about Black people (Devine, 1989) was supported for both the driving and the facial recognition scenarios. Research has demonstrated that people can perceive faces of ingroup and outgroup members in a biased way, and that this bias can manifest as a lack of empathy for the outgroup (Molenberghs & Louis, 2018). These differences in empathy may explain why, for the driving scenario, our results were not in the hypothesized direction. That is, although the participants were not explicitly asked to envision one of the people on the screen being struck by a self-driving car, they were still asked to decide which person would be struck in the instance of an accident. Thus, it is implied that the person they choose will be susceptible to some level of pain and the other will not. The fact that White participants spent significantly more time looking at the White faces compared to the

Black faces when given this scenario could be related to the expression of empathy to others in pain.

The term “empathy” refers to the ability to share and understand subjective states and feeling of others (Molenberghs & Louis, 2018). In a study conducted by Xu, Zuo, Wang and Han (2009), Chinese and Caucasian participants viewed video clips of Chinese and Caucasian people receiving either painful (i.e., needle prick) or non-painful (i.e., cotton swab) stimulation to the face. When participants viewed painful stimulation of outgroup faces there was more activation in the dorsal anterior cingulate cortex (dACC) and anterior insula (AI), areas known to respond when empathy for the pain of others is being expressed. However, when viewing outgroup faces in pain, no increased activation was observed in the dACC (Xu, Zuo, Wan, & Han, 2009).

Relatedly, research has asserted that conscious attention is required to process the emotions of others (Hofelich & Preston, 2010). The results of our study could indicate that participants spent more time looking at ingroup faces than outgroup faces due to the fact that they were better able to recognize and empathize with the emotions and pain that their decisions could cause to the ingroup members pictured on the screen.

For the facial recognition scenario, despite the fact that it was presented in a negative manner, it was hypothesized that participants would spend more time looking at the White faces in order to adequately compare the descriptors given to the ingroup face in front of them.

Previous studies have shown that stereotypically Black features are associated with crime and violence (Blair, Judd, & Chapleau, 2004). Thus, a current area of interest is whether or not stereotype-consistent categorization can influence memory and classification of Black faces (Kleider, Cavrak, & Knuycky, 2012). The results of this study could indicate that because outgroup members are likely to apply stereotypes to Black faces relatively quickly in a criminal

context, they do not need to spend as much time looking at the Black faces compared to White faces to make decisions. Most people are unaware of this implicit face categorization, but it has the potential to influence how Black faces are perceived and remembered, as well as how judgments are made about the individuals (Kleider, Cavrak, Knuycky, 2012).

Our second hypothesis that participants who had more outgroup contact over the course of their lifespan thus far would demonstrate less bias than those who have not experienced a significant amount of outgroup contact over the course of their lifespan thus far was not supported. One possible reason for this lack of an effect may be the fact that outgroup contact was relatively low in the current sample. This may at least be partially due to their attendance at a predominantly White university. That is, there may not have been enough variability to find significant correlations between contact and eye tracking patterns. Future research should examine a more heterogeneous population. A lack of statistical power may also be partially responsible for these results, given the small sample size.

The current study has several additional limitations. All of the data were collected from college aged students who are usually from higher socioeconomic backgrounds than the general population and have succeeded enough to be able to go to college. Yet there are some benefits associated with using a college aged sample: college is a time when young adults are contemplating their career path and may decide to get involved in the technology field, so their biases could end up influencing artificial intelligence programs. The U.S. Bureau of Labor Statistics (2022) reports that employment in computer and information technology occupation is predicted to grow 13 percent from 2020 to 2030, which is much faster than the average for all occupations.

Further analyses of the current data will examine the mouse tracking data. It is expected that, although participants spent more time looking at the White faces in the self-driving car and facial recognition scenarios, that they actually ended up selecting the Black faces more frequently. However, since there were not any significant results related to the hiring scenario with the eye tracking data, we do not expect the mouse tracking data to produce significant racial differences either. Future studies could select a different population from which to recruit participants to see if these results vary based on age of the participant or amount of experience working with programming artificial intelligence algorithms. Additionally, future research could use other racial groups (i.e. Asian or Hispanic) to see whether or not an outgroup member's perceived proximity to whiteness may affect the decisions that ingroup (White) participants make during each scenario. It would be incredibly interesting to see how implicit biases in this context may vary based on the racial differences across multiple groups.

Our findings support our primary hypothesis that participants would show racial differences in how they implicitly process two faces of people with different racial identities (i.e., Black and White) presented on a computer screen and that these differences would be based on racial stereotypes in our society conveying positive traits about White people and negative traits about Black people (Devine, 1989) for both the driving and the facial recognition scenarios. The current study did not find any significant correlation between level of outgroup contact and the expression of bias during the task. These results shed light on the way that implicit biases can impact the development and programming of AI.

References

- Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of afrocentric facial features in criminal sentencing. *Psychological Science*, 15(10), 674–679.
<https://doi.org/10.1111/j.0956-7976.2004.00739.x>
- Computer and information technology occupations: Occupational outlook handbook: : U.S. bureau of labor statistics.* (2022, April 18). U.S. Bureau of Labor Statistics.
<https://www.bls.gov/ooh/computer-and-information-technology/home.htm>
- Dilmegani, C. (2022, April 26). *Bias in AI: What it is, types, examples & 6 ways to fix it in 2022.* AIMultiple. <https://research.aimultiple.com/ai-bias/>
- Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82(1), 62–68.
<https://doi.org/10.1037/0022-3514.82.1.62>
- European Liberties Platform. (2021, May 18). *algorithmic bias: Why and how do computers make unfair decisions?* Liberties.Eu. <https://www.liberties.eu/en/stories/algorithmic-bias-17052021/43528>
- Face database.* (n.d.). Park Aging Mind Laboratory. <https://agingmind.utdallas.edu/download-stimuli/face-database/>

Hofelich, A. J., & Preston, S. D. (2012). The meaning in empathy: Distinguishing conceptual encoding from facial mimicry, trait empathy, and attention to emotion. *Cognition & Memory; Emotion*, 26(1), 119–128. <https://doi.org/10.1080/02699931.2011.559192>

Howard, A., & Borenstein, J. (2017). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, 24(5), 1521–1536. <https://doi.org/10.1007/s11948-017-9975-2>

Implicit bias (stanford encyclopedia of philosophy). (2019, July 31). Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/implicit-bias/>

Kleider, H. M., Cavrak, S. E., & Knuycky, L. R. (2012). Looking like a criminal: Stereotypical black facial features promote face source memory error. *Memory and Cognition*, 40(8), 1200–1213. <https://doi.org/10.3758/s13421-012-0229-x>

Kubota, J. T., Peiso, J., Marcum, K., & Cloutier, J. (2017). Intergroup contact throughout the lifespan modulates implicit racial biases across perceivers' racial group. *PLOS ONE*, 12(7), e0180440. <https://doi.org/10.1371/journal.pone.0180440>

Lin, Y. T., Hung, T. W., & Huang, L. T. L. (2020). Engineering equity: How AI can help reduce the harm of implicit bias. *Philosophy & Technology*, 34(S1), 65–90. <https://doi.org/10.1007/s13347-020-00406-7>

- Matzen, L., Haass, M., & McNamara, L. (2014). *Using eye tracking to assess cognitive biases* [Slides]. Sandia National Laboratories. <https://www.osti.gov/servlets/purl/1242146>
- Molenberghs, P., & Louis, W. R. (2018). Insights from fMRI studies into ingroup bias. *Frontiers in Psychology, 9*. <https://doi.org/10.3389/fpsyg.2018.01868>
- Nelson, G. S. (2019). Bias in artificial intelligence. *North Carolina Medical Journal, 80*(4), 220-222. <https://doi.org/10.18043/ncm.80.4.220>
- Najibi, A. (2020, October 26). *Racial discrimination in face recognition technology*. Science in the News. <https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/>
- Obermeyer, Z., Powers, B., Vogel, C., Mullainathan, S. (2020). Dissecting racial bias in an algorithm used to manage the health of populations. *Yearbook of Pediatric Endocrinology*. <https://doi.org/10.1530/ey.17.12.7>
- Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health, 9*(2). <https://doi.org/10.7189/jogh.09.020318>
- Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: Social

categorization and the process of intergroup bias. *Journal of Personality and Social Psychology*, 59(3), 475–486. <https://doi.org/10.1037/0022-3514.59.3.475>

Roff, H. M. (2022, March 9). *The folly of trolleys: Ethical challenges and autonomous vehicles*. Brookings. <https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/>

Silberg, J., & Manyika, J. (2019). Notes from the AI frontier: Tackling bias in AI (and in humans). *McKinsey Global Institute*.
<https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/MGI-Tackling-bias-in-AI-June-2019.pdf>

Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How mouse-tracking can advance social cognitive theory. *Trends in Cognitive Sciences*, 22(6), 531–543.
<https://doi.org/10.1016/j.tics.2018.03.012>

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, 19(6), 795–797.
<https://doi.org/10.1038/nn.4296>

Wong, H. K., Stephen, I. D., & Keeble, D. R. T. (2020). The Own-Race bias for face recognition in a multiracial society. *Frontiers in Psychology*, 11. <https://doi.org/10.3389/fpsyg.2020.00208>

Xu, X., Zuo, X., Wang, X., & Han, S. (2009). Do you feel my pain? Racial group membership modulates empathic neural responses. *Journal of Neuroscience*, 29(26), 8525–8529.
<https://doi.org/10.1523/jneurosci.2418-09.2009>

Appendix

Hiring Scenario:

“The company you work for as a hiring consultant has asked you to develop a system that will filter applicants based on the requirements of the position. The position that they are currently looking to fill is for an office manager. You will be presented with various faces and must decide which applicant seems better suited for the position.”

Facial Recognition Scenario:

“In order to crack down on crime, the local police department has decided to install cameras that utilize face scanning technology. They recently received a tip related to a bank robbery with a description of the perpetrator. The description was as follows: gender unknown, about 5’8”, brown or green eyes, and medium length hair. You will now be shown various faces. Please select the face that you think most likely fits the description. Your selections will be taken into consideration when programming the facial recognition software.”