The Pennsylvania State University

The Graduate School

IMPROVING THE PREDICTABILITY OF RACIAL PREFERENCE ATTITUDES: USING MACHINE LEARNING MODELS TO PREDICT CONCURRENT IMPLICIT-EXPLICIT BLACK-PREFERENCE ATTITUDES

A Thesis in

Information Sciences & Technology

by

Raphael A. Rodriguez

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2020

The thesis of Raphael A. Rodriguez was reviewed and approved by the following:

Frank Ritter
Professor of Information Sciences and Technology in the College of Information Sciences and Technology
Thesis Advisor

Daniel Susser
Assistant Professor of Information Sciences and Technology and Philosophy

Shomir Wilson
Assistant Professor of Information Sciences and Technology

Mary Beth Rosson
Director, Doctoral Program of the College of Information Sciences and Technology

ABSTRACT

The goal of this exploratory paper is to answer the plea for applying novel machine learning techniques to big data in social psychology by a) establishing a more predictable variable of attitudes toward Black people, and b) creating a single model that integrates five previously established stereotype and prejudicial measures. From the publicly available data from Project Implicit (projectimplicit.org), I used the Implicit Association Test (IAT) on Race from 2018—this dataset included a user's implicit preference for White or Black people, demographic information, and self-reported racial attitudes. As a result, the analysis suggests that demographic and self-reported racial attitudes are better able to predict concurrent implicit-explicit Black-preference attitudes (CIEBA) as compared to implicit or explicit attitudes separately (Avg. R^2, MSE: CIEBA = .365, 0.388; explicit-only = .324, 0.672; implicit-only = .110, 0.876). Secondly, self-reported questions that matched prior prejudicial measures (e.g., social distance, victim-blaming, egalitarianism, intergroup anxiety, blatant negative stereotypes) were grouped appropriately and were able to validly predict racial attitudes (R^2= .458, MSE = 0.325). This analysis also identified critical thresholds within the stereotype and prejudice measures at which point racial preferences for White people over Black people significantly accelerates.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my master's thesis advisor, Dr. Ritter, who gave me the intellectual freedom to pursue an exploratory analysis of a social issue and computational techniques neither of us were experts in. His guidance and open-door policy helped me development new computational analysis techniques in the social science field. I would also like to thank Dr. Susser and Dr. Wilson for agreeing to be committee members without much advanced notice. Despite my thesis not being fully encompassed in their areas of expertise, they have provided me with invaluable edits and thought-provoking questions. Lastly, I would like to thank the members of the Applied Cognitive Science lab for their editorial contributions in creating a polished thesis.

Chapter 1

Introduction

Undoubtedly everyone has implicit social biases for almost every situation regarding at least one other person; implicit both in terms of automatic, uncontrolled responses as well as all attitudes and behaviors that an individual does not self-report—possibly due to wanting self-reflection or fear of public judgment. The importance of measuring implicit biases is critical to understanding individual social attitudes; however, this is still only a partial assessment of the holistic attitude. As an example of this issue/problem/aspect of social cognition, the primary purpose of this research is thus to predict an individual's racial preference (explicit-only, implicit-only, or a combination of the two) for Black people or White people, from self-reported questions to determine if these questions more accurately reflect a person's self-reported racial attitude or a combination of implicit and explicit biases. Furthermore, the analysis employed machine learning algorithms to maximize the model's predictive power, as well as, codifying and comparing developed theories of prejudicial and stereotypical biases.

Machine Learning in Social Science

Yarkoni and Westfall (2017) described the goal of psychological science as the quest to understand human behavior—to explain the cognitive causations and predict future behaviors. Traditionally, these two goals were acknowledged to be equivalent; uncovering the psychological underpinnings of human behavior would lead to the accurate prediction of future actions. The authors, however, argue that these ends are statistically and pragmatically separate, "[i]t may well be that in many areas of psychology, scientists will ultimately have to choose between (a)

developing complex models that can accurately predict outcomes of interest but fail to respect known psychological or neurobiological constraints and (b) building simple models that appear theoretically elegant but have very limited capacity to predict actual human behavior" (Yarkoni & Westfall, 2017). An offered explanation for this inconsistency is the over-valuation of the statistical "goodness of fit" of models based on existing data—possible overfitting—without testing the models' predictive abilities on unclassified data. Yarkoni and Westfall (2017) thus suggest using machine learning models in psychology for their flexible applications and robust analyses.

Similarly, Harlow and Oswald (2016) strongly argue for the use of big data in psychology, as it has proven to be irreplaceable in the technological industry, marketing, consumption, politics, social media, and it is even starting to be accepted widely in some academic disciplines. "Big data science can be instrumental in collaboratively working to uncover and illuminate cogent and robust patterns in psychological data that directly or indirectly involve human behavior, cognition, and affect over time and within sociocultural systems" (Harlow and Oswald, 2016).

Measures of Prejudice and Stereotypes

Racial prejudice toward and stereotyping of Black Americans has been individually conceptualized and measured in a variety of ways. Created by Bogardus (1923), and later modified by Crandall (1991), the social distance scale is one earliest of these measures. This measure quantifies the desired separation from a stigmatized group, both direct contact and relational contacts. Parrillo and Donoghue (2005) found that overall mean social distance scores have decreased over time—the overall mean social distance score of their study in 2005 was 1.45 while in 1966 and 1977, the mean scores of those studies were 1.92 and 1.93, respectively (a

lower score indicating higher tolerance). These separate studies suggest there might be greater acceptance of minority groups, even the minority groups that have newly emerged in the United States.

Prior outgroup prejudice and stereotype research have found intergroup anxiety to be pernicious to intergroup relations (Stephan & Stephan, 1985). Stephan and Stephan (1985) discovered higher levels of intergroup anxiety are correlated "with low levels of contact with outgroup members, stereotyping of outgroup members, and assumed dissimilarity to outgroup members". Though, the level of intergroup anxiety differs depending on an individual's (a) knowledge of a given outgroup, (b) interaction with the outgroup, or (c) concern with public perception of prejudiced attitudes (Britt et al., 1996). Some of the negative stereotypes from intergroup anxiety may include higher estimates of the criminal rate of outgroup members, decreased support for policies favoring outgroup members, and higher rates of toxic behavioral intentions (Stephan, 2014).

As Katz and Hass (1988) discovered, the core values of American culture can create opposing affects towards outgroup members. The first value orientation is defined by a commitment to social justice and underdog support—egalitarianism—which promotes further communalism and equality. Even some individuals with high-prejudiced attitudes believed those prejudices should be tempered as a result of their moral obligations, rather than social pressure to appear non-prejudiced (Monteith & Walters, 1998). However, the Protestant work ethic values individualism and self-reliance, which leads to victim-blaming; yet, this blame does not evidence the broader cohesion of the community or strength of familial relations.

Lastly, the knowledge of existing negative stereotypes, as well as personal acceptance of those stereotypes, leads to blatant prejudices against the perceived inferior outgroup (Devine, 1989). Devine (1989) found that these blatant prejudices are so cognitively entrenched to reverse

these beliefs and attitudes "requires intention, attention, and time." Devine's findings were later supported by the results of Greenwald et al.'s (1998) Implicit Association Test.

Implicit Association Test and Prior Implicit Social Cognition Research

The Implicit Association Test (IAT) was developed by Greenwald et al. (1998) to measure an individual's implicit social cognition. During the test, participants are presented with two target categories (e.g., Black people and White people), they are then asked to match these target categories with differing attributes (e.g., positive and negative valence words) as quickly and accurately as possible. This technique is assumed to reveal the relative association strength of the target and attribute variables by measuring the separate response latencies of the association-compatible and association-incompatible assignments (Hoffman et al., 2005).

Much of the research involving implicit social cognition has involved measuring the correlation with and predictive validity of self-reported attitudes and observed behaviors (Greenwald et al., 2009; Cameron, Brown-Iannuzzi, & Payne, 2012; Kurdi et al., 2019). This research typically involves performing statistical analyses (e.g., regressions, ANOVA) on a single implicit social cognition measure (e.g., the IAT score) along with an individual's self-reported attitudes or behaviors—each individually investigated. This type of analysis is necessary to uncover the correlations and connections of these variables; however, neither measure on their own provide a comprehensive representation of an individual's social attitudes.

Chapter 2

Methods

The goal of this exploratory analysis is to use machine learning regression algorithms and big data to illustrate the importance of using more advanced computational techniques in social psychology by providing further insight into commonly studied phenomena. To this end, the user data comes from a large open-source dataset of implicit social cognition, the IAT from Project Implicit, and will be fed into a random forest regression and newly developed boosting algorithm from Microsoft.

Participants

The data used for this research was supplied by Project Implicit, a non-profit organization that measures and codifies implicit social cognition through the Implicit Association Test (IAT). Project Implicit publicly releases the data from their demonstration website (projectimplicit.org) of their various IAT tests—age, Arab-Muslim, Asian, disability, gender-career, gender-science, Native, presidents, race, religion, sexuality, skin-tone, weapons, and weight—which test for the implicit social bias toward a group (e.g., associations of women with the sciences or the arts). This exploratory research analyzed the Race IAT data from 2018, which contained 360,000 completed entries (less than 50% of the original 859,000 recorded entries in the file)—which included the implicit score, three self-reported Black attitude questions, demographic information, and one of the self-reported question groups.

This test is internationally available online, however, given the inherent complex association structure of Black people as a concept across, limiting the included users to a standard

demographic profile enhanced the model's accuracy. The user data was limited to residents of the United States who were between the ages of 18 and 72. After restricting the data set to those conditions, as well as removing incomplete or unclear entries, the finalized dataset contained about 117,000 entries.

In addition to implicit racial preference scores of each user (explained below), two other sets of information were analyzed, demographics and self-reported (explicit) answers to race-relevant Likert questions. The analyzed demographic information included: age, education level attained, political ideology, race, ethnicity, gender, religion, and religiosity.

The dataset also contained 31 self-reported questions groups, each group containing 3-23 questions (Appendix A), for a total of over 400 self-reported items—each of these question groups are mostly analyzed separately from each other. For ease of understanding, these question groups could be thought of as 31 separate studies included in a single dataset. In general, these questions ask users to report their attitudes toward Black people or broader racial prejudices, either directly (e.g., Do you feel warm, cold, or neither warm nor cold toward Black people?) or more subtly (e.g., I try to hide any negative prejudicial thoughts to avoid negative reactions from others.).

Variable Creation

There were two types of analysis performed. Random forest regressions were used to compare the correlation coefficients (r) and coefficients of determination (R^2) of the CIEBA, explicit-only, and implicit-only black racial attitudes. The Explainable Boosting Machine was used to interpret the effects each prejudice and stereotype variables have on the overall model.

Concurrent Implicit-Explicit Black-Preference Attitude (CIEBA)

The experimental dependent variable of the models was a combination of the IAT score (implicit) and three self-reported (explicit) race attitude preference questions—concurrent implicit-explicit Black-preference attitude (CIEBA). The self-reported racial preference attitudes and the IAT score were each separately used as the control dependent variables.

Implicit Score

The IAT score is a measure of the time latency difference between the two trials of associating target (i.e., White and Black people) and attribute (i.e., positive and negative) words. The IAT D score (Greenwald, Nosek, & Banaji, 2003)—the score used in this analysis and most other analyses—divides the difference between those two trials by the standard deviation of all the latencies of each assessment. The D score is already calculated and provided in the downloaded file; the D score generally ranges from -1.85 and 1.85. The mean of the IAT D score was 0.278, a positive value indicating an overall preference for White people. To be able to average the D score with the other explicit scores and improve the efficacy of the variable (Mohabeer, H., Soyjaudah, K. S., & Pavaday, N., 2011), the score was Z-score normalized.

Explicit Scores

Every user was asked the same three self-reported race attitude preference questions (explicit1, explicit2, explicit3), shown in Table 2-1. To integrate the self-reported variables, the difference between explicit2 and explicit3 was calculated to create explicit4—a single variable to compare preference among the two races. Then the explicit1 and explicit4 variables were Z-score normalized across the entire dataset to create explicit1_z and explicit4_z, respectively.

Table 2-1: The self-reported race attitude preference questions and possible answers; outside the other question groups and given to every user.

Explicit Questions	Possible Answers
Explicit1: Which best describes you?	1. "I strongly prefer African Americans to European Americans." 2. "I moderately prefer African Americans to European Americans." 3. "I slightly prefer African Americans to European Americans." 4. "I like European Americans and African Americans equally." 5. "I slightly prefer European Americans to African Americans." 6. "I moderately prefer European Americans to African Americans." 7. "I strongly prefer European Americans to African Americans."
Explicit2: Please rate how warm or cold you feel toward the following group – White people	0. "Extremely cold" 1. "Very cold" 2. "Moderately cold" 3. "Somewhat cold" 4. "Slightly cold" 5. "Neither warm nor cold" 6. "Slightly warm" 7. "Somewhat warm" 8. "Moderately warm" 9. "Very warm" 10. "Extremely warm"
Explicit3: Please rate how warm or cold you feel toward the following group – Black people	0. "Extremely cold" 1. "Very cold" 2. "Moderately cold" 3. "Somewhat cold" 4. "Slightly cold" 5. "Neither warm nor cold" 6. "Slightly warm" 7. "Somewhat warm" 8. "Moderately warm" 9. "Very warm" 10. "Extremely warm"

CIEBA

My new construct, CIEBA—the single more comprehensive and complexly descriptive Black attitude variable—is the mean of implicit1_z, explicit1_z, and explicit4_z. This derivation is shown in Eq 1 to 6.

(1) implicit1 = IAT raw score

(2) explicit1_z = explicit1 (normalized)

(3) explicit2_z = explicit2 (normalized)

(4) explicit3_z = explicit3 (normalized)

(5) explicit4_z = explicit2_z – explicit 3_z

(6) CIEBA = (implicit1_z + explicit1_z + explicit4_z)/3

Prejudice and Stereotype Measures

To analyze the individual effects of each prejudice and stereotype measures, a single variable for each of the five measures (i.e., social distance, victim-blaming, negative cultural stereotypes, egalitarianism, intergroup anxiety) were created from the most significant self-reported questions. To determine the self-reported questions that most expressly highlighted the contrast between those with strong anti-Black attitudes (SABA users)—those who significantly prefer White people over Black people—and the rest of the users (non-SABA), the means of each question were calculated. A user was considered to have strong anti-Black attitudes if their CIEBA score was at least one standard deviation above the average CIEBA score for the whole dataset (8,050 users, 6.9% of all users). The average score for each of the self-report questions (~400) was calculated for the SABA group and the non-SABA group; then, the effect sizes were calculated for each question. The questions with at least a medium effect score (Cohen's D =>

0.5) were kept—62 questions total. Finally, these questions were separated into five groups based on the prior literature of stereotype and prejudicial measures—social distance, victim-blaming, negative cultural stereotypes, egalitarianism, intergroup anxiety.

Coding and Formatting Variables

Additionally, it is essential to note that all of the stereotype and prejudice variables were coded and formatted, so higher values of the variable represented an increased preference for White people over Black people, see Table 2-2.

Table 2-2: The measures of each prejudice and stereotype variables and their representative values.

Variable	Variable Representation
Negative Cultural Stereotypes	Belief in negative cultural stereotypes of Black people; higher values indicate a further belief in those stereotypes
Social Distance	Desire for separation from Black people; higher values indicate heightened separation desires or practices
Egalitarianism	Belief Black people deserve equality and further support; decreased values represent more egalitarian beliefs
Intergroup Anxiety	Discomfort around Black people; higher values represent increased discomfort
Victim-blaming	Belief Black people caused their negative situations or need to be self-reliant in their responses to the situation; higher values indicate further support for this belief

Models

Two separate machine learning algorithms analyzed this dataset. The first algorithm, a random forest regression, is used to compare the coefficients of correlation (r) and determination (R^2) of the three dependent variables (i.e., CIEBA, explicitly-only, and implicit-only) for each of

the 31 self-reported question groups to identify the most predictive model. On the other hand, only one model was created with the Explainable Boosting Machine (EBM) algorithm, which incorporated the stereotype and prejudice variables with CIEBA and the dependent variable.

It also should be noted that when exploring which algorithms to finally use, I did also try to build a neural network model and considered building an agent based model. For practical reasons, I decided to not pursue either of these models. Having never built a neural model before, the time it would require to learn to build one, fully understand the results, and verify the veracity of those results seemed to be too costly. The IAT dataset is perfectly suited to build single agents—demographic and attribute values—however, agent based models require a quantifiably defined interactions between agents and the environment which would first require the type of analysis I describe in this paper. So the following regression models were chosen due to their ease of implementation, relatively clear interpretability of the results, and structure of the dataset.

Comparing CIEBA, Explicit-Only, Implicit-Only

For data-scientists, random forest regressions are a well-known, widely used, supervised learning algorithm. This machine learning program is what is referred to as a bootstrap aggregation, "bagging", learning method. Random forest runs many individual decision tree algorithms in parallel to each other—the decision trees train on random and replaced data samples—and aggregates the results of the decision trees to create a single model. This method reduces the problem of overfitting to the training dataset, thereby producing one of the most accurate predictive regression models as well as indicates the relative importance of each variable to the finalized model. However, the random forest algorithm is known as a "black-box" machine learning model as users cannot view the internal "decisions" made by the model, only the output is known (Liaw & Wiener, 2002)

The Random Forest Regressor, from the sklearn package in Python, was used to perform the following regression analysis. Three models were run on each of the separate 31 self-reported question groups—a total of 93 models. For example, for the ANES question group, the independent variables were the demographic information, ANES1, ANES2, ANES3, ANES4, ANES5, and ANES6; then the three dependent variables—the CIEBA score, the main explicit score (explicit1), and the single implicit score (implicit1), separately—were tested. This process was repeated for each of the question groups.

Modeling Prejudice and Stereotype Measure's Effect on CIEBA

The Explainable Booster Machine (EBM) with InterpretML, on the other hand, is a "white-box" machine learning algorithm, which is almost as accurate as a random forest algorithm, but with much greater interpretability of each variable. Historically, this has been the significant tradeoff of large models—they are either accurate or interpretable, but never both. Microsoft Research developed the Explainable Booster Machine in August 2019 (Nori et al., 2019).

An EBM model was created using the demographic information and the five stereotype and prejudicial variables—social distance, victim-blaming, negative cultural stereotypes, egalitarianism, intergroup anxiety—as the independent variables while the only dependent prediction variable was the CIEBA score.

Chapter 3

Results

In summary, based on the 93 models random forest models, the CIEBA measure is more predictable (R^2 = .365) from the demographic and self-reported questions of racial attitudes than either the explicit-only (R^2 = .324) or implicit-only (R^2 = .110) measures. Additionally, the EBM model depicted the prejudice and stereotype variables having a relatively significant impact on CIEBA scores. The figures of the prejudice and stereotype variables indicate that higher values of the variables correlated with increased CIEBA scores. However, the rate of acceleration differs among the variables, and for a few of the variables, the rate of acceleration dramatically increases once it surpasses a critical threshold.

Comparing CIEBA, Explicit-Only, Implicit-Only: Random Forest Regression

Table 3-1 displays the average correlation coefficients (r), the coefficients of determination (R^2), and mean squared errors (MSE) of the 31 models for each dependent variable. When comparing the R^2 score of the CIEBA model to the explicit- and implicit- only models, the CIEBA model was a 13% and 232% improvement, respectively. Similarly, the MSE of the CIEBA model was 42% more accurate than the explicit-only model and 56% more accurate than the implicit-only model.

Table **3-1**: The correlation coefficient (r), the coefficient of determination (R^2), and mean squared error (MSE) of the Random Forest Regression models with CIEBA, explicit-only, and implicit-only dependent variables.

Dependent Variable	r	R^2	MSE
CIEBA	.604	.365	0.388
Explicit	.569	.324	0.672
Implicit	.331	.110	0.876

Prejudice and Stereotype Variables: EBM

The Explainable Boosting Machine model supplies detailed information about the overall importance of each variable as well as their exact boosting score for the CIEBA value. Figure 3-1 shows the global importance—the mean absolute score—for the top 15 most critical variables to the model. Unsurprisingly, Table 3-2 shows whether or not a person identifies as Black or White is the most significant factor for preferring White people over Black people.

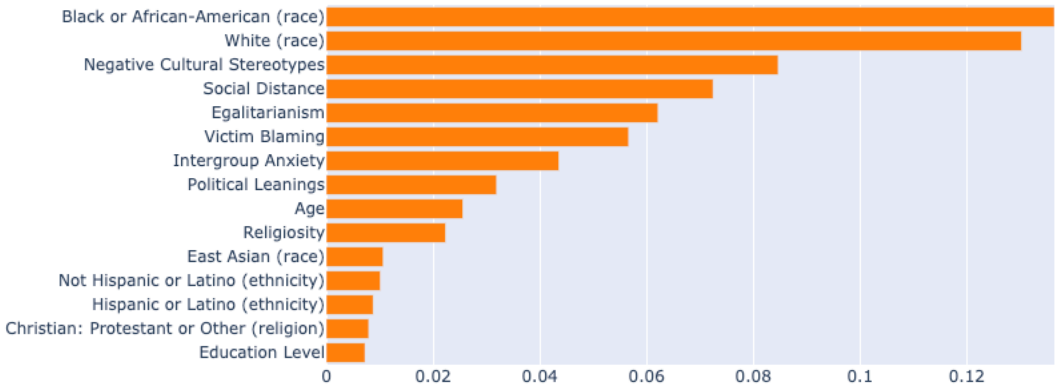

Figure 3-1: The overall importance each variable has on the final model calculated as the mean absolute boosting score. Racially identifying as Black or White are the factors that have the most significant impact on White over Black preferences, followed by the five prejudice and stereotype variables.

Table 3-2: Boosting scores for the top Race groups. Identifying as Black is the single largest factor for decreased White over Black racial preferences, followed by identifying as Non-White (i.e., American Indian/Alaskan Native, East Asian, South Asian, Native Hawaiian or Other Pacific Islander, Other or Unknown).

Race	Mean
Black or African American	-0.588
Non-White	-0.264
White	0.086
Non-Black	0.077

The predicted CIEBA value for a given user is calculated from the scores of the variables as answered by each user. For example, in Figure 3-2, the user self-reported a value of 1.17 on Negative Cultural Stereotypes—based on their answers to the relevant questions—which results in a 0.15 score boost to their CIEBA measure, as can be seen in Figure 3-7.

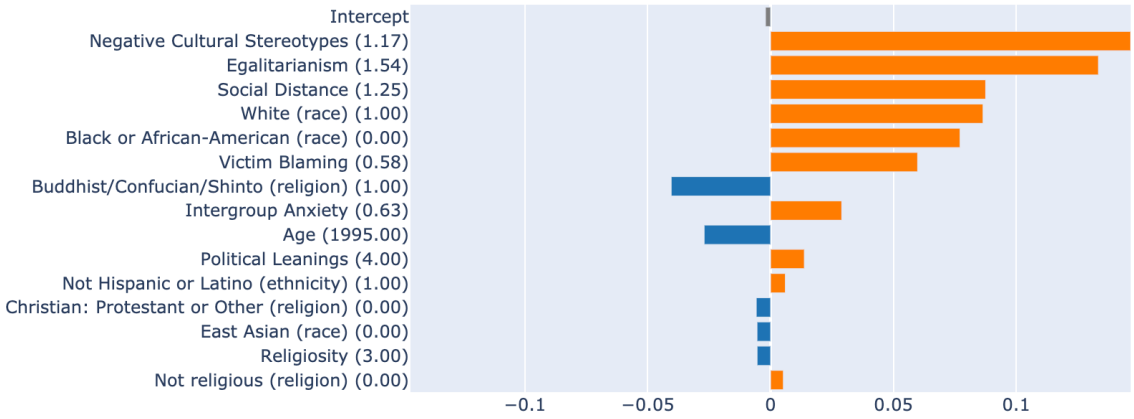

Figure **3-2**: The numerical boost each variable has on a person's CIEBA score—this was a random user selected from the training data. Actual CEIBA score is 0.56 and predicted is 0.57. The horizontal axis represents the variables' boosting valu,e and the numbers in parentheses next to the variable name are the self-reported values for each variable.

In other words, all else being equal, those with a Negative Cultural Stereotype score that are 1.17 standard deviations from the mean, on average, will have their CIEBA score increased by 0.15, an increased preference for Whites, Figure 3-2. Similarly, a user reporting to not be of White race results in a CIEBA score decrease of 0.264, Figure 3-3 and Table 3-2. The user represented in Figure 3-3 has a relatively higher racial preference for Black people over White people (CIEBA score: -0.20) than the user represented in Figure 3-2 (CIEBA score: 0.56).

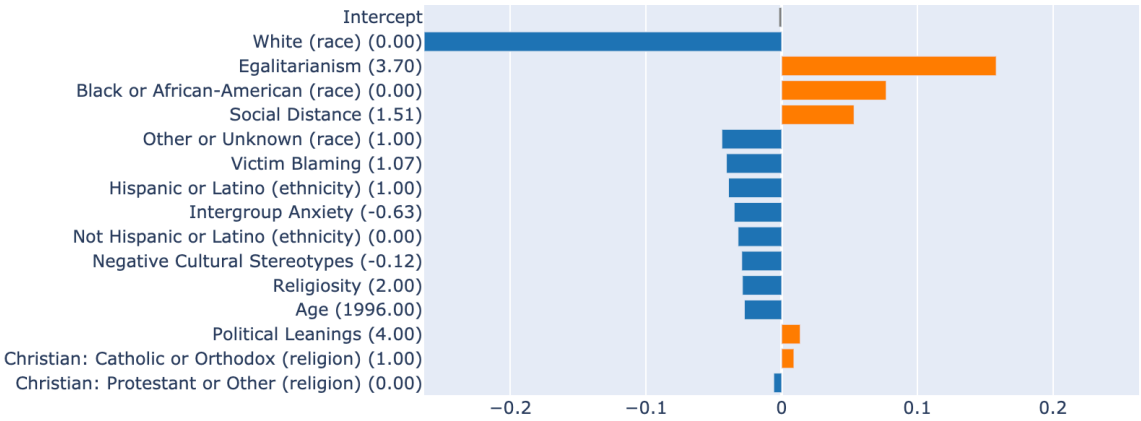

Figure 3-3: The numerical boost each variable has on a person's CIEBA score—this was a random user selected from the training data. Actual CEIBA score is -0.20 and predicted is -0.23. The horizontal axis represents the variables' boosting value, and the numbers in parentheses next to the variable name are the self-reported values for each variable.

Figures 3-4 – 3-8 show the corresponding CIEBA score adjustments to the values of each of the prejudice and stereotype variables. Each of those variables indicates higher CIEBA values as prejudice and stereotypes against Blacks increase; however, the intensity of the upward trajectory and points of substantial escalations differ. For example, an Intergroup Anxiety score of -0.63 standard deviations does not largely impact the CIEBA score (-0.04), while the CIEBA score is still only slightly boosted (0.12) if the Intergroup Anxiety score is ~3.5 standard deviations above the average (Figure3-7). However, the CIEBA values are much more affected by the Negative Cultural Stereotype scores. A Negative Cultural Stereotype score of -0.73 standard deviations reduces the CIEBA value by 0.11, but at 3.5 standard deviations, that variable increases the CIEBA value by 1.03, Figure 3-4.

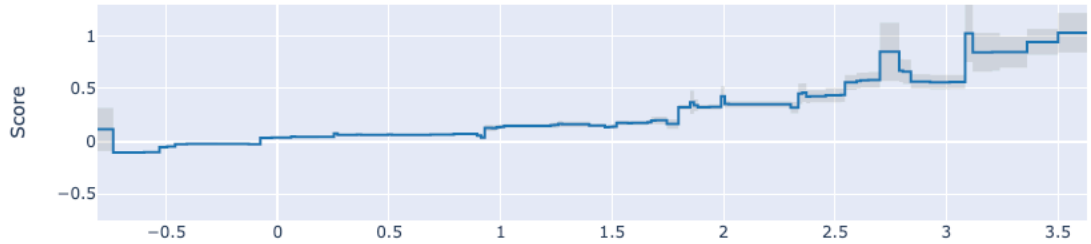

Figure 3-4: The boosting score for the entire range of values for the *negative cultural stereotype* variable. The boosting score moderately increases with an increase in the measure for most participants, but after a critical threshold (~1.80) is reached, the boosting score's growth quickly accelerates.

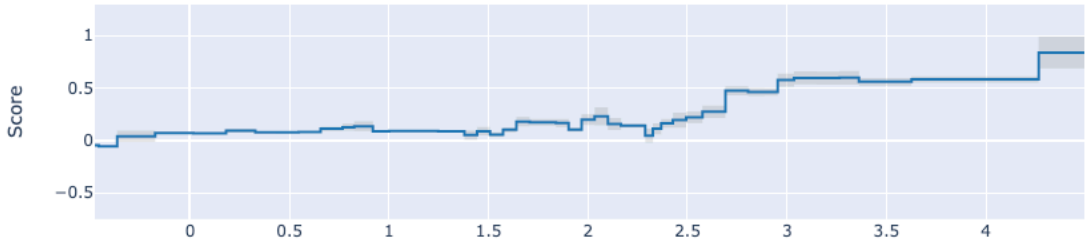

Figure 3-5: The boosting score for the entire range of values for the *social distance* variable. Similar to negative cultural stereotypes, this measure begins with a mild upward slope, but after a critical threshold (~2.31) is reached, the boosting score's increase growth accelerates.

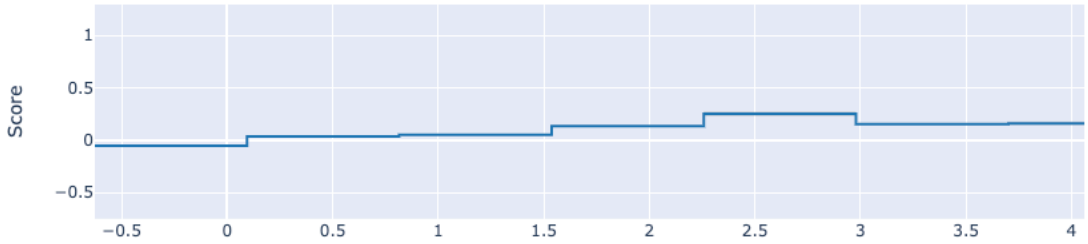

Figure 3-6: The boosting score for the entire range of values for the *egalitarianism* variable. The boosting score of this measure maintains a temperate positive slope of 0.04.

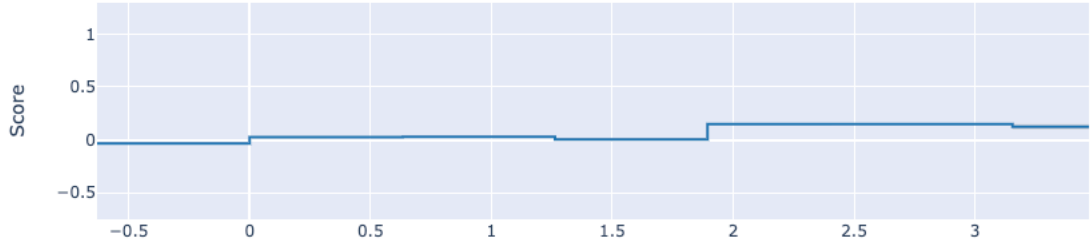

Figure 3-7: The boosting score for the entire range of values for the *intergroup anxiety* variable. The boosting score of this measure maintains a temperate positive slope of 0.04.

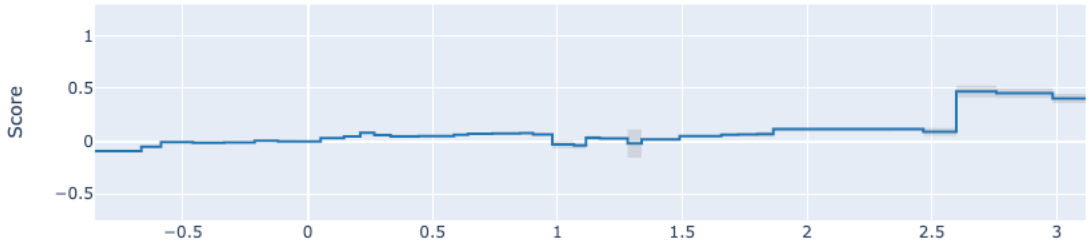

Figure 3-8: The boosting score for the entire range of values for the *victim-blaming* variable. The boosting score of this measure maintains a temperate positive slope of 0.05; however, at the critical threshold of ~2.53, there is a considerable spike in CIEBA scores.

Chapter 4

Discussion

Both of the machine learning algorithms provided significant insights into racial preference attitudes. The results of the random forest regression algorithm overwhelmingly supports the predictive superiority of using a combined implicit-explicit dependent variable. While the EBM and InterpretML algorithms provide a clear and interpretable plot of each variables' effect on racial preferences at all values, leading to the detection of a critical threshold at which the rate of preferring White people over Black people intensifies.

Comparing CIEBA, Explicit-Only, Implicit-Only

As expected, the explicit-only models out-performed the implicit-only models, likely for two reasons – a) the models included self-reported racial attitude responses and b) the legitimate criticisms of the IAT's ability to measure implicit biases accurately. The explicit variable was derived from the individual's blatant racial preferences, which on their own, are highly correlative with the other prejudicial and stereotype measures included as independent variables. Furthermore, the IAT, and implicit social cognitive measures, in general, have faced constant methodological and theoretical criticisms. Namely, the elastic susceptibility to socio-contextual factors (Barden et al., 2004) or the IAT is not an estimate of personal preferential attitudes but only the product of the awareness of those cultural stereotypes (Mitchell & Tetlock, 2017). These critiques, along with several others, have been previously addressed (see Jost, 2019); furthermore, the results obtained here add novel support to the value of the IAT.

Though the explicit-only models' superiority to the implicit-only models may be obvious, the reasons for the CIEBA models' dominance may be less apparent. The IAT's reliability faces

constant repudiation for low correlations with self-reported responses and demographic factors (Barden et al., 2004). However, the value of the IAT is evident when attempting to measure and predict a more complete assessment of a person's race preferential attitude. Including the implicit scores with the explicit scores has quantitatively shown to be a significant improvement in predictive racial bias modeling, see Table 3-1. These results support the notion that racial attitudes are more comprehensively represented by the concurrent implicit-explicit Black-preference attitude (CIEBA) construct. These results also seem to suggest that self-reporting biases, or other confounding factors, could be attenuated by including implicit measures, even though the implicit numbers by themselves are not wholly predictable.

Prejudice and Stereotype Variables

From the prejudice and stereotype variable figures (Figures 3-4 – 3-8), the five different variables can be categorized into two broad groups: critical threshold variables and consistent variables. First, the CIEBA boosting score of the negative cultural stereotypes, social distance, and victim-blaming variables steadily rises as the values of those variables increase, but once a critical threshold is reached, the rate of increase quickly accelerates. While the second group of variables—egalitarianism, intergroup anxiety—appears to have a relatively consistent boosting score, with mild inflations, as the values increases.

Critical Threshold Variables

As prior literature has evidenced, negative stereotypes or stigmatizations are highly correlative with social distance (Crandall, 1991). This relationship is visually represented by Figures 3-4 and 3-5; in both figures, the variables retain a relatively stable boosting score for

lower values, but the boosting score for each dramatically increases as people indicate higher negative stereotypes of or social distance from Black people. In the negative cultural stereotype figure, from the values -0.73 to 1.80 (boosting score range: -0.11 to 0.16), the average slope of the increasing boosting score is 0.10; yet for the values greater than 1.80 the average slope jumped to 0.48. Comparably, the average slope of the social distance values from -0.48 to 2.31 (boosting score range: -0.05 to 0.04) is 0.02; subsequently, the slope escalated to 0.36. As a result, these figures indicate that for most people, remembering that the values of the variables are standard deviations from the mean of the dataset, self-reported social distance and negative cultural stereotypes do not significantly affect their White over Black preference, as measured. However, once a critical threshold is reached (1.80 for negative cultural stereotypes and 2.31 for social distance), these two measures have a profound impact on racial preferences.

Even though these variables similarly affect the CIEBA score, it is still important to measure negative cultural stereotypes and social distance separately. Goff, Steele, and Davies (2008) found that racial stereotype threats and racial distancing were separate behaviors because there could be additional reasons for social distancing oneself from Black people other than blatant negative cultural stereotypes. As these two variables have the highest overall model importance of the stereotype and prejudice variables (Figure 3-3), as measured by the mean absolute score, continuing to measure and study these measures separately will improve the predictive accuracy of future models.

This critical threshold is also found, to a lesser extent, in victim-blaming. For a majority of the measure, self-reported victim-blaming has a moderate impact on CIEBA—values from -0.85 to 2.53 have a boosting score range of -.09 to 0.08 (slope: 0.05). However, once reaching that critical threshold (2.53), victim-blaming sharply becomes much more indicative of White over Black preferences.

Consistent Variables

Unlike the previous variables, self-reported intergroup anxiety and egalitarianism appear to lack a critical threshold for heightened White over Black preferences. The boosting score for the entirety of the values of the intergroup anxiety (boosting score range: -0.04 to 0.12) and egalitarianism (boosting score range: -0.05 to 0.16) variables remain at a relatively shallow positive slope, each at an average of 0.04. However, this is not an indication that these measures frivolously affect the CIEBA value; they still provide valuable insight into the individual contributions of separate prejudice and stereotype variables.

It also worth noting that although the egalitarian variable plot has an average positive slope, at ~3.00 there is a statistically significant dip in the boosting score—from 0.25 (95% confidence interval: 0.23 – 0.28) to 0.15 (95% confidence interval: 0.14 – 0.17). This type of statistically significant dip is not found in any of the other prejudice and stereotype variables. On the surface, this indicates that those who most strongly oppose social support for Black people actually prefer White people less than some of those who hold more pro-egalitarian values. However, an alternative hypothesis is that this phenomenon might be explained by a desire to over-compensate for their White preference due to social pressures to appear less prejudiced (see Butz & Plant, 2009). Further research and analysis are needed to discover the causation for this unintuitive.

Chapter 5

Conclusion

The results of the random forest models indicate that self-reported questions, regarding Black racial attitudes, more accurately reflect an individual's overall Black preference attitude, as opposed to only their self-reported attitudes. This finding is a critical distinction because answering multiple self-reported questions reveals more than self-reported biases; they more accurately predict contemporaneous implicit and explicit attitudes. Thus, when studying racial attitudes, and possibly other social biases, it would be more precise to concurrently measure implicit and explicit responses to indicate a truer bias.

The results of the Explainable Boosting Machine algorithm with InterpretML plotting provide several key insights. First, it directly provides the effect each variable value has on the predicted CIEBA value. This interpretability not only allows researchers to gain further insight into the effect a single participant's information has on their predicted scores, but this method also provides researchers with the ability to discover new inter-variable and intra-variable patterns within the entire model. For example, as discussed earlier, when reviewing the plots of the prejudice and stereotype variables, a quantifiable critical threshold emerges.

Second, the plots of each of these measures may provide additional support for these theories in social psychology, or they could potentially explain some of the anomalies as well. Additionally, this model and research may help future researchers, particularly those in the implicit social cognition field, identify the variables that have the most impact on racial attitudes or which measures to focus on if resources are limited. This underscores the potential impact of identifying critical thresholds for prejudice and stereotype measures in practical applications. When creating and measuring the success of social interventions, critical thresholds could serve

as important benchmarks; such as identifying intervention populations, intervention techniques, and maximum barriers. Moreover, these thresholds could be measured across locations and time to identify potential factors that affect the threshold limits and following accelerated pace.

Lastly, these algorithms give further credence to the increasing demand for using advanced machine learning techniques in psychology. These methods allow researchers to acquire new cognizance of long-standing psychological theories or potentially to establish additional nuances to the theories. However, machine learning regressions are just one example of advanced computational techniques that could benefit the social science fields; there are numerous advanced artificial intelligence (AI) methods that should be used by researchers in this field. For example, the algorithms governing the AI agents in games or extensive simulations—or, more specifically, the algorithms determining the social interactions of those agents—could be used to simulate social interactions based on real-world big data. Nevertheless, it is important to note that there is no single computational method which is best suited for all data structures or research fields. It is possible that using a different algorithm with this dataset would have yielded results which were not novel or a vast improvement from previously established techniques. Thus, it is vital for the social sciences to continue to incorporate advanced computational methods and document the successes and failures of each of those methods.

This study does have some limitations, however. First, there is a growing dispute of the IAT's validity to measure implicit biases (see Jost, 2019). Most of this debate centers around the limited implicit biases predictability, which is reflected by the low R and R^2 values of the implicit-only model (Table 3-1). However, this research is partially in response to that critique, even if the IAT score is mostly unreliable, when combined with the explicit measures, they add greater accuracy to the models. Lastly, given the novelty of the EBM algorithms, the code used and the models generated should be validated by additional research.

Looking forward, there are still many advanced computational analyses that can analyze IAT, or other implicit social cognition, data. There are also further subsets of the data to examine; this research only focused on one year and one country of the Race IAT data, but this data has been collected since the early 2000s and in most international countries. The self-reported questions can be further analyzed individually or combined differently to test other established theories in social psychology. Also, potentially the CIEBA score could be calculated slightly differently—change the weights of the implicit and explicit measures. Furthermore, additional models should be built for all the other IATs, at least 14 unique publicly available tests, which each have many years of collected data. However, a powerful potential for this field is combining this data with other large datasets to gain further understanding into not only predicting racial attitudes but some of the underlying systemic causes of social inequalities. Potentially incorporating data on government spending on social welfare, capitalism indicators, indices of democracy, or general indicators of social inequalities could help identify the systems and policies which correlate with social inequality

References

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual Moderation of Racial Bias: The Impact of Social Roles on Controlled and Automatically Activated Attitudes. *Journal of Personality and Social Psychology*, 87(1), 5–22.

Bogardus, E. S. (1923). *Immigration and race attitudes*. Boston: D.C. Heath.

Britt, T. W., Bonieci, K. A., Vescio, T. K., Biernat, M., & Brown, L. M. (1996). Intergroup anxiety: A person× situation approach. *Personality and Social Psychology Bulletin*, 22(11), 1177-1188.

Butz, D. A., & Plant, E. A. (2009). Prejudice control and interracial relations: The role of motivation to respond without prejudice. *Journal of Personality*, 77(5), 1311-1342.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential Priming Measures of Implicit Social Cognition: A Meta-Analysis of Associations With Behavior and Explicit Attitudes. *Personality and Social Psychology Review,* 16(4), 330–350.

Crandall, C.S. (1991), Multiple stigma and AIDS: Illness stigma and attitudes toward homosexuals and IV drug users in AIDS-related stigmatization. *J. Community. Appl. Soc. Psychol.*, 1: 165-172.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56(1), 5.

Goff, P. A., Steele, C. M., & Davies, P. G. (2008). The space between us: Stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology*, 94(1), 91-107.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21(4), 447.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, 31(10), 1369-1385.

Jost, J. T. (2019). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science*, 28(1), 10-19.

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structures. *Journal of Personality and Social Psychology*, 55(6), 893.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., ... & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, 74(5), 569.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.

Mitchell, G., & Tetlock, P. E. (2017). Popularity as a poor proxy for utility. In S.O Lilienfeld & I.D. Waldman. *Psychological science under scrutiny: recent challenges and proposed solutions* (pp. 164-195). West Sussex: Wiley & Sons.

Mohabeer, H., Soyjaudah, K. S., & Pavaday, N. (2011, August). Enhancing the Performance of Neural Network Classifiers using Selected Biometric Features. *In Proc. 5th International Conference on Sensor Technologies and Applications, French Riviera, Nice/Saint Laurent du Var, France*.

Monteith, M. J., & Walters, G. L. (1998). Egalitarianism, Moral Obligation, and Prejudice-Related Personal Standards. *Personality and Social Psychology Bulletin*, 24(2), 186–199.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223*.

Parrillo, V. N., & Donoghue, C. (2005). Updating the Bogardus social distance studies: A new national survey. *The Social Science Journal*, 42(2), 257-271.

Stephan, W. G. (2014). Intergroup Anxiety: Theory, Research, and Practice. *Personality and Social Psychology Review*, 18(3), 239–255.

Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of Social Issues*, 41(3), 157-175.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122.

Appendix

Number of Questions per Group

1. ANES – 6 questions
2. ATB – 20 questions
3. ATW – 20 questions
4. BRSF – 6 questions
5. CAB – 6 questions
6. CAW – 6 questions
7. CC – 9 questions
8. EFP – 12 questions
9. GIA – 12 questions
10. GSRACE – 22 questions
11. GSSOPP – 9 questions
12. IA – 11 questions
13. IMSEMS – 10 questions
14. MC – 10 questions
15. MCPR – 17 questions
16. MR – 7 questions
17. NR – 7 questions
18. OP – 6 questions
19. PAAQ – 20 questions

20. PGC – 5 questions
21. PINDEX – 10 questions
22. RAB – 20 questions
23. RAW – 21 questions
24. RR – 6 questions
25. RSM – 6 questions
26. RWAF – 20 questions
27. SBP – 20 questions
28. SDOF – 16 questions
29. SR – 8 questions
30. TP – 3 questions
31. UO – 20 questions