Using YouTube Advertisements to Reduce Racial Bias
Research Proposal

November 27th, 2020
Robert Hosbach, Mikayla Pugel, Sarah Xie

Overview

Due to the internet age and usage of social media, society is becoming increasingly polarized. Some of this polarization is natural. Cognitive bias—a known limitation of being human— is a systematic error in the way people interpret and process information (Bagalini, 2020). This effect can cause a variety of biases[1] of varying degrees, and as a society we have always lived with these limitations. However, the last 10 years has observed an increase in this polarization ("The Shift", 2020). Many experts theorize that this increase is attributed to the biased effects of using data to tailor what individuals see and think (Johnson, 2020). As a result, people have a harder time understanding others who may not see the world the same way, as the differences in worldviews tend to be extreme. The overall goal with this research is to provide necessary tools to the client, The Center for Humane Technology[2], to decrease this pattern. Herein we provide a research proposal to study the change in human bias over time, with the goal to create a more open-minded society.

Bias is strongest when assessing individuals that bear a different physical appearance. Humans naturally lean towards people who look like them and act like them, which is a common groupthink phenomena (Murphy, n.d.). These groups provide a sense of comfort and belonging that everyone strives for. However, diversity is important in order to have a sustainable society (O'Boyle, 2020). This report will detail a proposed quantitative research study, including the data collection, sampling, statistical methods, intervention, potential risks, and the final deliverables to understand if diversity in advertising reduces racial bias. We will partner with Youtube to obtain the participants for the study and place them randomly into a control and treatment group. We will then test for their current implicit bias towards different races using a version of the Harvard Implicit Bias Test in order to measure any changes in this bias after the study's intervention ("Project Implicit", 2020).

Research Question

Do participants that demonstrate partiality for people of their own race become less biased when they are shown positive advertisements featuring people of races other than their own?

Study Design

This study will take a quantitative approach to the data that is gathered through the time-series experiment described here.

To begin, we will work with YouTube to poll a sample of their users for volunteers to participate in this study. YouTube-users will be asked if they are willing to participate in a five-week virtual study on implicit biases that involves very minimal work on their end. We will include a link to a

[1] A bias is an inclination for or against something or someone.
[2] https://www.humanetech.com/

webpage with a longer description of the study, and the webpage will also include other necessary links the participants will use throughout the study. Those who sign up to participate will be asked to sign a disclaimer that if they experience extreme reactions to the content they're shown throughout the study, the responsibility will not fall on the researchers' shoulders. The disclaimer will also contain information about the study and state the researchers' commitment to safe and ethical practices. After signing the disclaimer, participants will need to take a short pulse survey to capture important demographic indicators (age, sex, race, and geographic location).

Once enough volunteers have been gathered to fit the desired sample size, participants will be separated into control and experimental groups by random assignment. Participants will each be given a participation number (to hide their names or other identifying factors) from 1 to 1000, and a random number generator will pick 500 numbers from this range. Those 500 numbers will be assigned to the experimental group, and participants corresponding to the other 500 numbers will form the control group. Test and control groups are crucial to understanding whether there is a causal relationship between the advertisements that participants are shown and the change in their bias score.

After the control and experimental groups are assigned, all participants will be given an implicit bias test (IBT) modeled after the Harvard Implicit Bias Test (Project Implicit, 2011) to determine what their current state of bias is against individuals of a different race. The test will present images of people that are the same race as the participant, juxtaposed against images of people that are not of the same race as the participant, and the participant will be required to associate words with each image. The key metric measured here will be the "bias score", a number generated by the test that quantifies how biased the participant is against races that are not their own based on the words they associate with each image shown throughout the test.

This test is the first of two, and it will act as the baseline bias that each individual will be compared to after the study is over. After being given a week to take the first test, participants will undergo a three-week treatment, during which they will use YouTube as normal but receive differing advertisements according to their biases and test/control group placement. This treatment is described in more detail in the "Intervention" section of this document. Finally, at the end of this three-week period, participants in both test and control groups will be given a week to retake the IBT they were given at the beginning to track the changes in the level of their biases.

Once the experiment has ended, the results will be analyzed through a Wilcoxon Rank-Sum Test to determine whether the participants' bias scores changed significantly from the beginning of the experiment to the end. The change in the experimental group's bias scores will be compared against the control group's bias scores in order to remove any noise from the results. More detail is provided in the "Statistical Methods" section.

Data

The data used for this research would be collected throughout the study and would capture how the bias of the study subjects evolved after the treatment. The final dataset would include the participant's "bias score", which will be measured on a scale and can be captured through a test designed by the experimenters (modeled after the Harvard IBT, mentioned in the "Study Design" section). It will also include a timestamp for when that score was given and a description of the different advertisements the participant underwent after their first attempt at the IBT and before their final attempt.

Sample

The sample will consist of individuals who are 1) between 18 and 65 (inclusive) years of age, 2) U.S. citizens, 3) identified by Google as heavy and consistent viewers of YouTube media who consistently allow advertisements to play through, and who 4) agree to participate in the study. We will enlist 1,000 participants for this study, with 500 participants randomly assigned to each of the control and experimental groups (as described in the Study Design section).

Intervention

For this study we will have two experimental groups. The control group will continue to receive the standard YouTube advertisements on their social media accounts based on YouTube's native algorithms. The experimental group, in contrast, will receive targeted YouTube advertisements that are positive[3] and that feature actors that are of a different race than the participant in an effort to counteract the participant's racial bias. The advertisements will escalate in intensity, starting with some individuals of the participant's race mixed in with the actors. By the last week of the experiment, participants will receive advertisements with no actors of the participants' race (*i.e.*, participants will only see advertisements featuring actors of races that are different from the participant's). Throughout the study, it is crucial that the advertisements shown to participants in the experimental group will still align with the participants' hobbies and interests (similar to the advertisements the participants in the control group will receive based on YouTube's native advertising algorithm).

Statistical Methods

We will perform a Wilcoxon Rank-Sum Test for this research study. The assumptions for this test include that the two groups, test and control, are independent groups, and that the two populations have equal variance. We have chosen this test for two reasons. First, we can not assume that the data is normally distributed and therefore can not use a Student's *t*-test. Second, we have chosen a non-parametric test as we are not trying to estimate a parameter, but find evidence of a distribution shift caused by the intervention. This test will be completed with a confidence level of 95%. The null hypothesis and alternative hypothesis for this test are stated below.

H_0 : *There is no difference in bias scores from before and after the treatment.*
H_a : *The after-treatment bias score is lower than the before-treatment bias score.*

[3] Neutral or positive advertisements are those that do not directly attack a product, person, idea, etc.

We have chosen a one-tailed test because we want to know if the experimental treatment leads to a decrease in racial bias. After completing a Wilcoxon Rank-Sum Test, if the p-value is larger than the chosen limit (0.05) then we will fail to reject the null hypothesis and if the p-value is smaller than the chosen limit, then we will reject the null hypothesis. The rejection or failed rejection of the null hypothesis will help us determine the practical significance of this research design.

Potential Risks

We have identified multiple potential risks for this study:

- **Effect longevity**. Even if the results of this study suggest that racial bias can be decreased over the experimental period, the experiment may not have lasting effects. To determine the effect longevity, the Center for Humane Technology could conduct a follow-up with study participants months, or even years, after the conclusion of the study.
- **Short treatment period**. The treatment will last for three weeks, which may not be long enough to observe any effect. This could be addressed by increasing the study duration, which would allow for a longer treatment period.
- **Ecological validity**. The sample of participants involved in this study will not be representative of the population that uses social media (or even just YouTube). This study's participants only include adults from the United States who are consistent YouTube users that do not commonly skip YouTube advertisements and who are willing to participate. This risk can only be mitigated by selecting a large random sample from the entire sphere of social media users.
- **Self-selection bias**. Persons who meet the sampling criteria will be asked if they are willing to participate in this study. The persons who agree to participate in the study may be more inclined to believe racial bias exists and have a greater desire to minimize their own racial bias compared to those who do not agree to participate in the study. If a participant's desire to change their existing racial bias impacts the success of shifting that participan's bias (due, for example, to the social desirability effect), the study results may be skewed. To avoid self-selection bias, participants should be chosen randomly and given the treatment regardless of their consent, which is unethical and not recommended.
- **Attrition**. Participants may explicitly drop out of the study or stop taking the interim IBTs during the course of the study. To mitigate this risk, we will provide study participants with 1 month of free YouTube TV service for each IBT they complete. Participants who complete all of the IBTs during the course of the study will therefore earn four free months of YouTube TV service. The data from participants who do not complete the study will not be included in the final analysis; but, the attrition rate for the study will be provided in the final report.
- **Random assignment**. The separation of control and experimental groups relies on the use of a random number generator, the efficacy of which has been questioned (Wichmann & Hill, 1982; Marsaglia et al., 1990). The purpose of using a random number generator (or any other method of random assignment) is to protect the study and its results from researchers' potential selection biases. But, if the random number generator

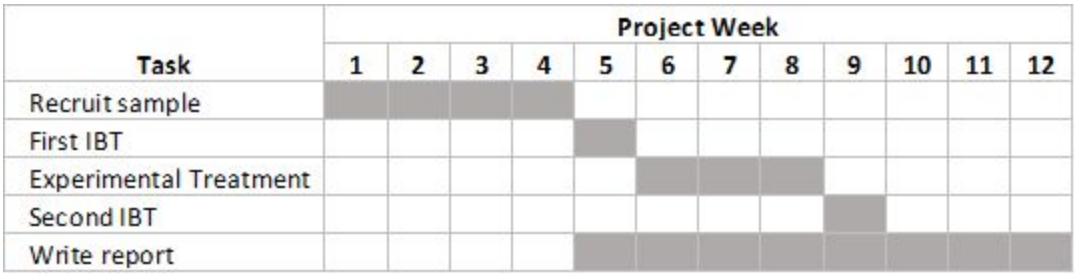
does not produce statistically random numbers, this could introduce another bias into the study.

Deliverables

Prior to the commencing the experimental portion of the study, we expect to spend up to one month recruiting our sample. We will begin this process by meeting with YouTube representatives to discuss their past surveys and studies and how long they expect the recruitment process to take. Then, participants will be given one week to complete the first IBT. The experimental portion of this study will be conducted over three weeks. Following the experimental phase, participants will be given one week to complete their final IBT. We expect to analyze the results and submit a final report within three weeks to a month of the experiment ending. The final report will be structured as follows:

- Executive Summary
- Introduction
- Methods
    - Sample Frame
    - Statistical Analysis
- Results
- Conclusions, Applications, and Future Work
- References
- Appendices
    - Details on Statistical Analyses

Overall, the research study will take three months to complete. The Gantt chart below shows the expected timeline for each task.



References

Bagalini, A. (2020, August 19). 3 cognitive biases perpetuating racism at work - and how to overcome them. Retrieved December 01, 2020, from https://www.weforum.org/agenda/2020/08/cognitive-bias-unconscious-racism-moral-licensing/

Johnson, S. L. (2020, October 28). Opinion | Facebook serves as an echo chamber, especially for conservatives. Blame its algorithm. Retrieved December 01, 2020, from https://www.washingtonpost.com/opinions/2020/10/26/facebook-algorithm-conservative-liberal-extremes/

Marsaglia, G., Zaman, A., & Tsang, W. W. (1990). Toward a universal random number generator. Statistics & Probability Letters, 9(1), 35-39.

Murphy, K. (n.d.). The Psychology of Polarization And How We Can Overcome Our Prejudices. Retrieved December 01, 2020, from https://www.bridgealliance.us/the_psychology_of_polarization_and_how_we_can_overcome_our_prejudices

O'Boyle, T. (2020, June 25). 5 Reasons Why Diversity is Important in the 21st Century. Retrieved December 01, 2020, from https://ampglobalyouth.org/2020/06/20/5-reasons-diversity-important-21st-century/

Project Implicit. (2011). Race IAT. Retrieved December 02, 2020, from https://implicit.harvard.edu/implicit/takeatest.html.

The Shift in the American Public's Political Values. (2020, July 02). Retrieved December 01, 2020, from https://www.pewresearch.org/politics/interactives/political-polarization-1994-2017/

Wichmann, B. A., & Hill, I. D. (1982). Algorithm AS 183: An efficient and portable pseudo-random number generator. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *31*(2), 188-190.

Statements of Contribution

- Mikayla Pugel: Wrote the overview, research question, and statistical methods sections. Helped edit the overall report.
- Robert Hosbach: Wrote the sample, intervention, and deliverables sections; contributed to the potential risks section; and edited the overall report.
- Sarah Xie: Wrote the Study Design and Data sections of the report, and contributed to the Potential Risks section. Helped structure & edit the overall report.