The Bias in Bias

Anubhav Arora
Massachusetts Institute of Technology

Nava Haghighi
Massachusetts Institute of Technology

Abstract - Bias in Bias is an interactive visualization on explicit and implicit racial bias in the United States. This visualization uses the Harvard Implicit Association Test (*IAT*) racial bias dataset. The primary goal of this visualization is to de-stigmatize the notion of bias by demonstrating that biases are inherently a human quality, educate and demonstrate the difference between implicit and explicit biases, and allow a exploration of the dataset at a deep level minimizing possible strengthening of the user's existing biases.

The visualization walks the user through a narrative that is split into three parts, each communicating a different piece of the narrative; 1. Biases are inherent to being a human and all humans are biased 2. Average numbers do not tell the full story 3. Biases are a result of many external factors, factors we may not have had any control over especially during development.

The final exploratory visualization consists of an interactive map which allows exploration of the dataset on a county level to emphasize the nuances in the data (the state average would lose such nuance), a histogram showing the distribution of bias for different groups, and an interactive scatterplot displaying the stories of individuals in the selected dataset.

Keywords -- Racial Bias, IAT, Data Visualization
Index Terms -- Racial Bias, IAT, Data Visualization

1 INTRODUCTION

Implicit bias is different from known bias in a way that it exists subconsciously. This means that a person might not think of themself as prejudiced and consciously not have any biases, but unknowingly demonstrates biased behavior. This scenario can be especially dangerous because revealing the unknown biases can question the person's fundamental beliefs about themself, making them defensive. One reason for this is because the society attaches negative connotations to any types of bias and most people do not realize that bias is inherent to human nature. Although certain types of biases are bad, looking for and recognizing one's implicit biases can help mitigate the negative impacts of bias.

This project aims to highlight the idea that bias is not inherently bad, and that everyone is biased in some way. The truth is that the human brain is noticing patterns all the time and making associations and generalizations. This is not to say that all our intentions are pure. What we hope to convey through this visualization is the fact that people have implicit biases they are not aware of. One step towards activating our recognition and acceptance of biases is de-stigmatizing bias and distinguishing between explicit bias and implicit bias. It is natural and human to have biases and it is important to understand that being biased does not simply make you a bad person. Additionally, biases are a byproduct of factors such as environment, culture, income level, gender, and education, many of which are not under our control all the time. No one is born biased. However, only when we recognize and acknowledge our implicit bias we can start to counteract its influence on our actions and decisions.

2 RELATED WORK

Research on Race IAT

The Implicit Association Test (IAT) started in 2002, and since then has garnered a lot of attention, especially in academia. It has been cited over 4000 times on Google Scholar and is widely used as an implicit measure in psychology. Current body of work related to implicit bias falls into three primary categories; (i) Understanding the interpretation of IAT scores and applicability (Blanton, Jaccard, Strauts, & Tetlock, 2015; Hahn, Judd, Hirsh, & Blair, 2014), (ii) Understanding external factors that might influence implicit bias (Hehman, Flake, & Calanchini, 2018), and (iii) Visualization of IAT results to inform policy changes.

Data Visualization and Bias

While there is ample research on the biases and distortions that certain visualizations propagate (Pandey, Rall, Satterthwaite, Nov, & Bertini, 2015; Sarikaya, Gleicher, & Szafir, 2018, Danielle S, 2018), their applicability to racial bias data is limited. Nonetheless, these research works offer interesting starting points. Sarikaya et al posit in their work that the way data is summarized can influence user in a significant way. Pandey et al summarize that user can be deceived or biased by a visualization either at the chart level or at the message level. While chart level deception occurs at the visual encoding level and depends on the user's familiarity with the chart, message level deception occurs at the message interpretation level, and may lead to creating biased beliefs about the message and its components. This is especially applicable to the existing visualizations using the IAT dataset where the data is used to support the author's claims and therefore is visualized in a static and possibly subjective manner, without allowing the user to explore and extract their own insights from the dataset.

3 METHODS

Our objective and exploration through our visualization was two-fold (i) how data can be used to start a conversation around a socially challenging topic such as racial bias, and (ii) how to do that in a way that doesn't enforce existing biases. We divided our project into 4 stages: Cleaning and Exploring data, Exploring and Developing visualization mock-ups, Preparing the final data set, and Creating the final data visualization.

3.1 Cleaning and Exploring the IAT dataset

The IAT project has publicly available data on 14 different kinds of implicit biases. Each implicit bias--ranging from age, race, gender, religion, etc.-- has its own online test and associated data set. Most studies have been collecting data online for around 10 years. The race dataset in particular has data since 2002; The full data set has over 7M rows and over 800 variables. Most variables refer to questions that test takers are asked during the test. These include demographic based questions as well as certain explicit questions that help determine the explicit bias of a user. Other variables are calculated during the test (such as time take to answer a particular question) and are used to calculate the final IAT score. Since the test is more than 15 years old, the biggest challenge with data clean-up was changing questions from year to year. It seems that a lot of questions (both demographic and explicit) are tested each year, and based on user response, they are either continued or discontinued. Another challenge with the dataset was many missing values--most of the demographic and explicit-based questions are not "required" and the test taker can skip a question if she/he doesn't desire to answer it.

3.2 Design explorations and creating visualization mock-ups

Several design directions were initially explored. One direction was *temporal analysis* of the data and visualizing the race IAT scores over time for different demographic and geography groups. The second direction was *spatial analysis* and visualizing the race IAT scores for different states/counties in the US. The third kind of visualization that we explored was *co-relationship* of the race IAT score with different demographic variables such gender, age, race, education level. We found that although these visualizations were allowing us to uncover some interesting insights from the data (such as on an average male individuals are more biased towards white people than females), they were doing it in a way that was strengthening our biases. We realized that looking at averages at a state, county, or demographic levels alone was the primary reason for this. We therefore decided to drill down to each individual level and show other external factors (such as whether an individual's parents are biased or not, or whether the individual was discriminated as a child or not, or what is the racial diversity of the neighbourhood of the individual, etc.) that might influence bias.

Another goal we had set to achieve was on communicating a narrative that informed the user of the human nature of biases, and defined the difference between implicit and explicit bias. One direction for this aspect of the project was to develop a participatory installation through which the user would receive a bias "badge" that visualized the user's implicit bias using the different IAT test results. This badge also communicated their explicit biases, demonstrating that many people while ideologically and consciously agree with unbiased behavior, may subconsciously have biased behavior. We ultimately decided to communicate this narrative through walking people through and purposefully asking them to generalize and show biased behavior, then revealed their biases to them through this interaction.

3.3 Preparing the final dataset

Three datasets were merged to create the final data set. The demographic information of the participants (including their gender, race, age, education, citizenship etc,) and the individual race bias scores were pulled from the IAT data set. The county-level information on diversity index, % of people with college degrees, and median income levels were pulled from the opportunity index dataset. Finally, latest FIPS county codes from US census data were used to allow us to map it visually. Data cleaning, wrangling, and final subsetting were achieved by using packages and libraries in R.

3.4 Creating the final data visualization

The final visualization consists of three sections; section 1 reveals the user's biases to them through an interactive exploration of select demographics from the dataset. This section may show the user how their biases are wrong, but most importantly demonstrates that bias is natural and human, we all generalize facts with limited number of data references all the time, and that everyone is biased.

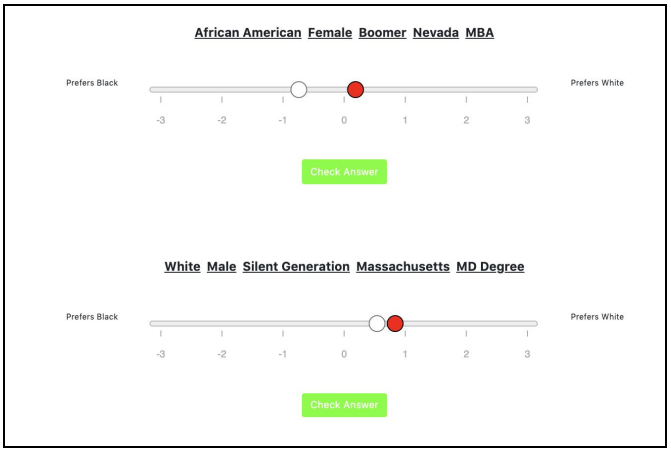

Figure 1: Bias is human, everyone is biased. Red dot marks the actual data and white dot shows the user's guessed average.

The second section emphasized that while we tend to think that people are representative of their demographic average, that may not always be true. In this section we ask the user to log their own racial bias and compare it to their demographic group's average to show that they may not be the average of their group. Additionally, in this section we introduce the difference between explicit bias and implicit bias and show that the answer they just logged most likely shows their explicit bias.

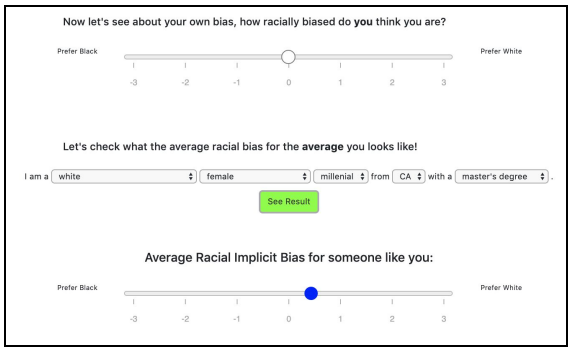

Figure 2: You are not the average you.

Section 3 consists of a map that visualizes both implicit and explicit bias data at a county level. The user can select subsets of the data based on demographics. The selected subset is also shown in a histogram that shows the distribution for both explicit and implicit bias in the United States, as well as plot explicit against implicit bias in a scatterplot.

The tooltip on the county shows deeper information about each county such as demographic information and county diversity. The tooltip on the scatterplot shows deeper personal information on each individual such as the bias in their parents or whether they have been discriminated against. This is to help the user understand some of the external factors beyond demographics that can influence bias.

To visualize the final interactive visualization, we explored multiple available tools and libraries. Our selection criteria for the tools was based on the speed and efficiency, ease of use, and adaptability and flexibility of the tool. We explored map libraries such as Leaflet.js and Mapbox GL JS but decided to use D3.js for the final visualization because it was the most flexible and could be easily integrated into the other aspects of the visualization.

For filtering and processing the data, due to the large size of our dataset we decided to combine Crossfilter.js with DC.js due to its speed of performance. However eventually we had to steer away from using Crossfilter.js due to DC.js's limited map libraries. Finally we explored Vega and Vega-Lite which ranked high on ease of use but performed slower than expected due to the size of our dataset.
Our final prototype was implemented in D3.js V4.

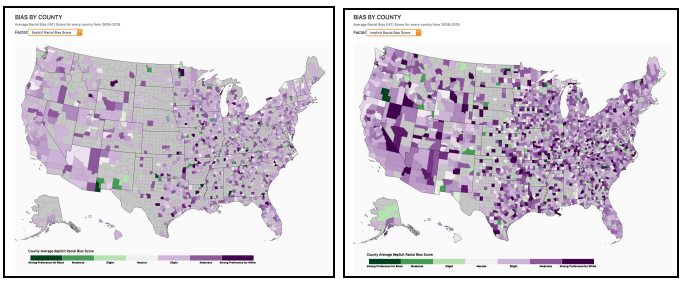

Figure 3: Difference in implicit and explicit bias for selected group

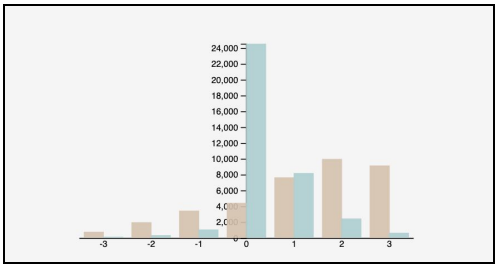

Figure 4: Distribution in implicit & explicit bias for selected group

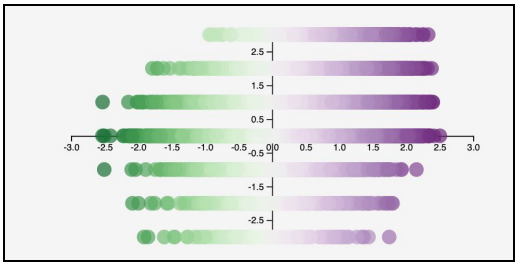

Figure 5: Scatter of all people in the group and biases

4 RESULTS

We anecdotally tested our visualization with real users for *interactivity*, *engagement*, *comprehensibility*, and *biased-ness*. We define interactivity in terms of the ease with which users were able to explore the visualization. We define engagement in terms of the degree of exploration that the user indulged in. We define comprehensibility in terms of the level of user understanding of our data and insights from it. Finally, we define biased-ness as the degree to which the user was left biased by the visualization. This to us was one of the most important metrics since our big idea was to visualize racial bias in a way that wouldn't enforce existing biases.

For interactivity, we believe that our visualization was satisfactory. We observed that the user was easily able to interact with the visualization without being prompted. However, there were certain characteristics that the user didn't interact with. These included zooming into the map and the zooming and moving around the scatter plot.

For engagement, our visualization performed satisfactorily as well. We observed the users were engaging well with the estimation questions at the beginning and then later with the map (specifically to observe the average bias of their own county). The engagement with other charts vis-a-vis the bar and scatter was not as much.

For comprehensibility, we believe that our visualization performed well. Users were able to understand the overall story and objective of the visualization without many prompts. The difference between implicit and explicit bias was well understood as well as their distribution for different groups.

For biased-ness, our main objective, we believe that our visualization performed really well. We had several follow-up conversations with our users who appreciated the fact that the exploration metrics were not limited to the usual suspects of person

demographics. Additionally, people became aware of their personal biases through the interactions and made remarks on how the results matched their expectations or not, and also how those expectations uncovered some of their personal biases. We were also able to use the visualization to start difficult conversations around racial bias, and people didn't feel as if they were being attacked or judged.

5 DISCUSSION

This visualization demonstrated an example of visualizing a dataset that is likely to strengthen existing biases. We demonstrated a direction for visualizing this dataset that can make the user aware of their own biases, but not get dissuaded by that information. Instead, the visualization allows the user to explore external factors beyond numerical and demographic averages to appreciate the complexities around the concept of racial bias.

Our visualization was tested against user interactivity, engagement, comprehensibility, and biased-ness. We realized that there are several aspects that the visualization could be further improved upon, specifically on interactivity and engagement. We believe that adding visual markers such as + and - could aid the user's intuition towards the zooming capability. We also believe that both reducing the amount of text and increasing the size of the text on the screen would improve user engagement. A few other limitations are discussed below.

5.1 Limitations

- Although we used a 'zoomable' scatter plot divided into four quadrants (filtered by the selected group) to show individual level bias scores, we weren't satisfied with the visualization. The points were too crowded, and we realized that it was hard to easily gather useful insights.
- The data in the filters was not ordered, this made it difficult for certain users to find their state.
- We believe that highlighting interesting stories and data points in the charts could have allowed users to better navigate through the different charts.
- Limited number of data points for certain counties and selected groups could bias the average IAT score one way or the other.
- Zooming and tooltip is an effective technique, however brushing tool could have made exploration more intuitive.

6 FUTURE WORK

6.1 Improving the visualization

As discussed in the limitations section, we believe that our visualization could be further improved. Cross-interaction and communication between different charts (especially the geographic map and scatter) would make exploration easier. Testing the visualization with more users is another step towards improvement.

6.2 Using our work to de-stigmatize the conversation around bias

Our visualization shows that racial biases are present and in fact their existence is inevitable. However, our visualization goes a step further and highlights that racial bias might be influenced by external factors. Future research work could further explore other external factors that might influence one's racial bias.

6.3 Building more informed and conscious visualizations that don't further existing biases

While building our visualization, we decided to not limit ourselves to the usual demographic information. We believe this allowed our visualization to communicate the complexity around racial bias instead of relying on mere averages. Future work could further explore ways to make visualizations more data neutral.

REFERENCES

1. Blanton, H., Jaccard, J., Strauts, E., & Tetlock, P. E. (2015). "Toward a meaningful metric of implicit prejudice": Correction to Blanton et al. (2015). Journal of Applied Psychology, 100(5), 1482–1482. https://doi.org/10.1037/a0039215
2. Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. Journal of Experimental Psychology: General, 143(3), 1369–1392. https://doi.org/10.1037/a0035028
3. Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents. Social Psychological and Personality Science, 9(4), 393–401. https://doi.org/10.1177/1948550617711229
4. Pandey, A. V., Rall, K., Satterthwaite, M. L., Nov, O., & Bertini, E. (2015). NELLCO Legal Scholarship Repository How Deceptive are Deceptive Visualizations?: An Empirical Analysis of Common Distortion Techniques. Retrieved from http://lsr.nellco.org/nyu_plltwp%0Ahttp://lsr.nellco.org/nyu_plltwp/504
5. Sarikaya, A., Gleicher, M., & Szafir, D. A. (2018). Design Factors for Summary Visualization in Visual Analytics. Computer Graphics Forum, 37(3), 145–156. https://doi.org/10.1111/cgf.13408