Allegations of Algorithmic Bias: Investigation and Meta-Analysis

Authors: gilesh@ (Privacy Analyst)
mcphillips@ (Public Policy)
vinaygoel@ (Privacy Analyst)
woodruff@ (Security & Privacy UX)
Counsel: cohenn@, kcooke@, wdevries@ (Privacy Legal)
Sponsors: lyou@ (Director of Privacy)
marisajimenez@ (Public Policy)
Last updated: September 2016
Go link: go/allegations-of-algorithmic-bias

Executive Summary
Introduction
Methodology
Example Allegations
Common Themes
Recommendations
Acknowledgments

Executive Summary

During the first half of 2016, the authors investigated several external claims of algorithmic bias in Google products to understand the nature of these claims and Google's organizational response to them. Most claims had previously been investigated to some degree, so we reviewed relevant documentation and met with stakeholders to learn more, and in one case collaborated with members of the Trust and Safety User Advocacy team who were currently leading an ads experiment.

Our investigation revealed key opportunities for Google to improve its handling of algorithmic bias: (1) a **coherent cross-product position**; (2) **systematic testing;** and (3) **improved external reporting mechanisms**.

Introduction

Algorithmic bias is an increasingly prominent topic in public policy, the press, and academic circles. Google is a frequent target of criticism in this debate, and also cares deeply about the ethics of its algorithms. Therefore, it is worthwhile to revisit Google's approach and ensure that it has excellent mechanisms in place to identify and address potential algorithmic bias.

Methodology

To select the allegations, we surveyed a number of stakeholders such as Communications, Public Policy, and Trust and Safety, as well as conducted an informal search of external publications and

press. From the list we assembled, we chose several allegations that had diverse characteristics. For example, we chose allegations about several different product areas, with different affected populations.

Our team then took an incident-response approach to investigating these allegations. We created a post-mortem-style incident template, which included information such as details of the allegation, the timeline, results of any internal testing, and what went well and what could have gone better. For each case, we met with key stakeholders who represented different perspectives (e.g., the product team, Product Policy, Communications) to learn what had happened. We also reviewed relevant documents where they existed. For each of the allegations, we completed a report based on our template. For the Ad Fisher allegation, we collaborated with members of the Trust and Safety User Advocacy team who were conducting an investigation in order to uncover the root cause of the behavior.

After completing the reports, we conducted a meta-analysis to identify common themes.

Example Allegations

In this section, we describe four of the cases we investigated, as illustrative examples.

Instant Checkmate: Latanya Sweeney published an article in 2013 claiming that Instant Checkmate search ads suggesting an arrest record tend to appear with black-associated names, and ads for public records from several companies tend to appear with black-associated names. An internal investigation conducted after the fact was not able to verify this finding; it is not clear whether it would have been reproducible at the time Sweeney reported observing it. If it had been reproducible, there are a number of potential explanations (e.g., Instant Checkmate listing names of individuals of many ethnicities as keywords, but happened to win the bidding war a disproportionately high number of times for black-associated names; or Instant Checkmate listed black-associated names as keywords as Sweeney suggests).

Sunlight Study: Columbia researchers published a paper in 2015 claiming that Gmail content that included terms related to health, race, religious affiliation or religious interest, sexual orientation, or difficult financial situation was associated with targeted advertisements for those topics. The researchers claim this violates a Google statement that it will not target based on these categories of sensitive information. The researchers' system cannot assign intention of either advertisers or Google for the targeting found. Ad product teams were unable to reproduce this claim. However, ad product teams could envision ways in which this type of targeting could occur. Because of that potential, changes were made in the Gmail ad targeting process.

Ad Fisher Study: CMU published a study in 2015 with experiment-based observations, arguing that Google's ad serving system perpetuates gender bias on the basis of two campaigns that were found to target high salary jobs at male users on the Times of India website. The effect was highly sensitive to the ads from this particular service, and the same effect was not reproduced in several other experiments by the same authors. The cause was found to be higher CPA (cost per conversion) for female users for one campaign (notably with a higher CTR for female users) and advertiser targeting to male-only users for the other.

Gorillas Mislabeling: In June 2015, a web developer posted on Twitter that Google Photos had tagged an image showing him and a friend at a concert with the label, Gorillas. “Of all terms, of all the derogatory terms to use,” Alciné said later, “that one came up.” According to a post mortem, Google executives noticed Alciné’s tweets within one hour. A Googler reached out through Twitter for permission to access the user’s photos; the issue was identified and resolved. In the short term, the Photos team stopped suggesting the “Gorillas” tile in the Explore page and stopped showing Search results for queries relevant to gorillas; in the long term, the team pledged to investigate the image annotation models that generate gorilla label false positives and work on the image annotation pipeline and quality evaluation processes.

	Instant Checkmate	*Sunlight Study*	*Ad Fisher Study*	*Gorillas Mislabelling*
Affected Party	Individuals with names associated with black individuals	Health, Race, Religion, Sexual Orientation, Finances	Women	Black users
Property	Search Ads	Gmail Ads	Display Ads	Photos
Reporting Date	2013	2015	2015	2015
Reporting Mechanism	Article	Academic Paper	Academic Paper	Twitter
Veracity	Unknown	Unknown	True	True
Root Cause	Unknown (many possible causes if true)	Unknown (many possible causes if true)	higher eCPM; advertiser targeting to an all-male remarketing list (was Unknown until recent investigation)	Difficult photo to classify; user testing did not flag sensitivity of labeling humans as gorillas
External Comms	Reactive statement (specific) + background points	Reactive statement (specific) + background points	Reactive statement (general) + background points	Reactive statement (specific) + background points
Product/Policy Change	Unsure (no longer serve ads based on proper names in some countries, but not sure what prompted change)	Yes (changes to Gmail ad targeting process)	No change (non-interference)	Yes (restrictions on the use of “gorillas” in the product)
Trust & Safety Involved	Yes	Yes	Yes	Yes

Table 1. Summary Table for Examples

Common Themes

The following themes emerged from our investigations:

1. **Prior Testing:** Limited or no testing had been done to identify or prevent these issues before they were reported.
2. **Reproducibility:** The alleged behavior was difficult to reproduce for the allegations related to ads, but easy to reproduce for the other allegations. For example, it is difficult to reproduce specific effects after the fact because the specific ad may have changed, and the set of all ads currently in the auction has changed.
3. **Reporting Mechanism:** Reporters went through highly visible channels such as Twitter or academic publication, in all cases. Some of these channels such as academic publications had long time delays, which made the issues harder to investigate.
4. **Responsiveness:** Google was generally highly responsive in investigating and addressing the claims, both in terms of policy/product changes and in terms of external communications.
5. **Product or Policy Change:** Product policy and/or product behavior were changed in response to several allegations.
6. **Internal Alignment:** Product policy, product behavior, and public relations goals were not fully aligned, in some cases.
7. **Central Stakeholders:** In most cases, Trust and Safety (previously PQO) was heavily involved in the resolution of the allegation, and Trust and Safety Product Policy had relevant product policies in place (although in some cases modification or interpretation was necessary). In most cases, Communications was also involved.

Recommendations

Based on our observations, the following three issues are particularly worthy of further attention:

1. **Coherent Cross-Product Position.** Many of the policies and strategies currently in place to limit algorithmic bias appear to have evolved organically and locally to product teams. Further, product behavior, product policy, public relations goals, and public policy goals are not always fully aligned. While a uniform policy is unlikely to be appropriate, a more consistent and coherent set of cross-product policies would be extremely valuable.
2. **Systematic Testing.** Google would benefit from a systematic approach to reducing and testing for algorithmic bias. This approach would cover issues such as general mechanisms to automatically flag potential issues (with the recognition that this is a significant technical challenge, especially for complex systems such as ads), and diverse samples for training and testing data sets.
3. **Improved External Reporting Mechanisms.** External reporters are currently using highly visible mechanisms to report issues. A direct reporting mechanism for algorithmic bias issues may result in Google receiving more information about these issues in a more timely fashion, improving Google's ability to reproduce the instance, respond and also decrease external visibility.

Acknowledgments

Many thanks to jmcdonnell@ and jnalven@ of the Trust and Safety's User Advocacy Group for leading an experiment to investigate the claims made in the Ad Fisher study.

We are grateful to the following for their valuable help and insights: afaville@, aschan@, aschou@, askim@, awmcdiarmid@, haihongz@, jakara@, jmcdonnell@, jnalven@, mattdougherty@, mfalgoust@, nkim@, pchiu@, pchurch@, om@, opruzan@, recode@, robmahini@, rongge@, sdholland@, shas@, teresitap@, woojink@, xbao@.