Changing the narrative on bias in advertising

Research findings from IBM Watson Advertising on using artificial intelligence to detect and mitigate unwanted bias in advertising technology

IBM **Watson Advertising**

Understanding bias in advertising technology

This year, over 4.6 billion people will use the internet to find information, interact with friends and loved ones, read news and be entertained[1]. Those online experiences and the infrastructure that supports them, specifically digital advertising, are increasingly dependent on artificial intelligence (AI) and machine learning (ML). These advanced technologies help increase efficiency and accuracy, help users find and discover content, and provide relevant products and services.

Machine learning and artificial intelligence are also helping publishers and content providers to understand, innovate upon, and monetize the data they collect around digital consumption. Digital advertising, a key component of any organization's monetization strategies, is currently amidst a massive industry-wide transformation and will need to be heavily dependent on AI and ML as Google and Apple eliminate traditional identifiers.

While AI and ML have quickly assimilated into advertising technology and are making a significant impact, topics like bias, privacy and transparency dominate the discourse around advertising and marketing, and curiosity and regulation around these topics are growing. The industry increasingly relies on platforms that automatically segment and select audiences, deliver offers, and optimize creative—more critical decisions are being made by machines. The Federal Trade Commission has put forth positions on how companies should consider building and employing AI and ML while developing additional trusted AI guidance[2]. Chief marketing officers (CMOs) are concerned about how these technologies consume data, make decisions, and whether they are putting their relationships with consumers at risk.

CMOs are asking questions: What causes the system to decide to do one thing over another? Could the machine learning algorithm react to some unseen signal like gender or age? How do I trust the machine's decisions? Does the data that I collect meet my standards for addressing bias?

These aren't new concepts or points of concern. But the possibility for technological bias to exist and its ability to scale continues to be amplified by exponential growth in data and our industry's growing dependency on automation. Technological bias can occur when a human cognitive bias or biases in the training data are| unknowingly encoded into the system and distributed at scale. These unintended biases may become systemic issues that are often difficult to detect, especially in the complex interaction of data and signals within the digital advertising ecosystem.

"Machine learning and artificial intelligence are helping publishers and content providers understand, innovate upon, and monetize the vast amount of data they collect.

Meanwhile, The Federal Trade Commission has put forth positions on how companies should consider building and employing AI and ML while developing additional trusted AI guidance[2]."

It is a complex problem for a complicated industry that is already dealing with massive change, and therefore, governance can be challenging. Biases are ingrained in how humans think about and process information. Consider that the advertising industry has operated on the premises of past learnings and successes for 40 plus years. Though biases can find their way into the technology that is being built, the biases are put there by human thoughts, assumptions and judgments. While CMOs may be asking the relevant questions about biases, most organizations aren't leaning into the development of company-wide policies for the ethical use of data. A recent study by Media Post found that less than half of brands plan to develop policies for the ethical use of data, even though the potential for liabilities and negative consumer impact grows[3].

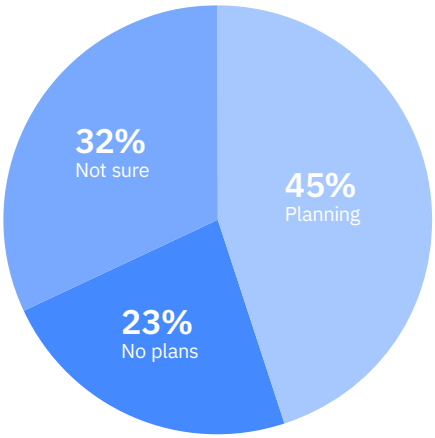
Consider a QSR (quick service or fast-food restaurant) marketer that is running a national ad campaign. While the campaign is broadly targeted, there is a possibility for the bid-optimization algorithms that assess when, where, and to whom the campaign is served to systemically overcompensate. The models may begin seeing engagement in largely lower-income areas in major metropolitan centers and begin to optimize performance from this observation. This might cause individuals in lower-income neighborhoods to experience an amplification of their lack of access to diverse food options. And on the other end of the spectrum, perfectly viable audience members in other areas, likely with a location nearby too, will see less and less of the ad campaign. The impacts from this type of unwanted systemic bias and the possible amplification could greatly harm brands with their campaign spend not delivering on the intended reach, and potentially a negative reaction from their audience. Possibly worse though is the social and personal impact it could have on the individuals and communities—an unwitting and unintended affirmation of their circumstance.

If a brand could keep a scenario like this from happening to a human, or their own brand, shouldn't they?

Research for trust

In addition to these headwinds, our cultural and societal standards are evolving for the better. Diversity, equity, and inclusion practices have shifted to the forefront of business decisions. Every organization must assume responsibility for building approaches and processes that put consumer trust at the center. With consumer confidence hanging in the balance, businesses must understand the technology that is being employed may have consequences. The importance of trust is evident where brands and consumers interact: digital advertising and marketing. While marketers must transform their messaging, positioning, and brand voice strategies to support diversity, equity, and inclusion initiatives, they will also need to work harder to understand the impact of the technologies they employ.

% developing policies for ethical use of consumer data[3]



Source: Advertiser Perceptions.
Base = 300 ad execs interviewed Sep 1-15, 2021

For consumers to trust the technology advertising and marketing employs, the advertising industry must develop and adopt better practices. Technology companies are investing in building better tooling for the development of AI systems. IBM Research has been developing a series of tools and policies to ensure that AI systems are built fairly and robustly, with algorithms that are easily explainable, while accounting for and aligning to the values and choices of consumers.

Bias research in advertising

In late summer 2021, IBM Watson Advertising initiated a research effort to understand how bias might impact digital advertising technology. Building the right team to do this research was especially important for this project's success. Demographic variety and experience were critical in forming the research team, as was the inclusion of advertising industry veterans and participants new to the practice. The team's makeup featured engineers, strategists, data scientists, researchers, product managers, and designers, which were essential to avoid group-based anchoring, bandwagon, and blind-spot biases about how the industry works, and a mixture of perspectives to help interpret observations.

The hypothesis

Biases are present in digital advertising data and algorithms, and those biases can be detected and mitigated using AI.

AI Fairness 360

The research team leveraged a toolkit called AI Fairness 360 (AIF360), employing open-source tools and building proprietary approaches. AIF360, developed by IBM Research, lets developers use state-of-the-art algorithms to help identify unintended biases within the machine learning workflow and mitigate uncovered biases.

This was the first time these tools were used in the digital advertising context. A digital advertising campaign is complex, and leverages different advertising technologies across various partners and channels. To fully understand the implications of bias in a campaign, the team needed to gather advertising data from across the different functions and pull it into one cohesive dataset. The team then worked with IBM Research to develop industry-first bias detection and mitigation methodologies modeled for the complexity that advertising creates and some of the limited events that often happen within a campaign. For example, the portion of users who click through an ad is only representative of a fraction of the overall targeted audience.

Exploring campaign data from The Ad Council

IBM Watson Advertising partnered with The Ad Council to understand the possible presence of unintended bias in the algorithms and data from one of the message testing efforts for their COVID-19 vaccine education campaign in partnership with the COVID Collaborative. The "It's Up to You" campaign ran throughout last year and focused on reducing vaccine hesitancy by encouraging people to get the answers they need to make an informed decision and protect themselves and their loved ones. The Ad Council utilized IBM Watson Advertising Accelerator as the predictive dynamic creative optimization technology to test one of their creative concepts with an audience based on designated market areas (DMA) and age. This test ran on The Weather Channel properties and across programmatic channels.

The research team utilized the collective data from this one flight consisting of 10 million impressions and over 108 different creative variations delivered via predictive dynamic creative optimization. After acquiring the campaign data, the team transformed them into one cohesive data set derived from log files across the advertising technology stack. This effort provided a transparent and seamless stack of data for post-processing experimentation.

Post-processing methods were used to run tests and experiments on data that has already completed its life cycle so as not to cause any adverse impacts on currently active campaigns through experimentation. While the research outcomes outlined below are based on post-processing, there is a potential to utilize these tools to impact a campaign in real time.

The research team used a combination of metrics and algorithms to identify groups within this sub-campaign data that were either systemically advantaged or disadvantaged. For example, Multi-Dimensional Subset Scan is an algorithm that detects subgroups in the audience that are anomalous relative to a predetermined target. The team used this approach to surface anomalous audience groups for additional analysis. To illustrate, the target was click-through rate on the predictive creative variant within the Ad Council campaign. Disparate impact ratio is defined as the ratio of the rate of favorable outcomes for the one group to the rate of favorable results for the other group, the two groups predetermined by the evaluator or surfaced by some other method like the Multi-Dimensional Subset Scan. When this ratio is observed to be less than 1, the first group is considered disadvantaged compared to the second group. Similarly, if this ratio is much larger than 1, the first group is considered to be at a relative advantage. Depending on the scenario, this ratio can vary widely, say from a value close to 0 to a value much larger than 1. These numbers represent the data or algorithms' bias towards or against specific groups within an audience and could be due to bias in the training data or some inherent unintended bias in the way the algorithms are designed and optimized.

"The research team used a combination of metrics and algorithms to identify groups within the campaign data that were either systemically advantaged or disadvantaged."

While exploring groups that surfaced during testing, the team sought to understand the characteristics appearing in disadvantaged and advantaged groups. Through experimentation with the segmentation data, the team established 250 Proxy Characteristics that might ladder up to individual characteristics. For example, home ownership status, or value of home owned might reference back to more deterministic economic characteristics like income. One of the questions this fostered was "Does the machine learning algorithm observe proxy signals at run-time and can they surface within biased groups?"

"The metrics suggested that bias mitigation would foster broader, fairer exposure to messaging predicted to convert."

The Ad Council's test was set up to target liberal and conservative-leaning DMAs. These DMAs were divided into age groups. While one might think that any bias that appeared as a result of bias in the algorithms would be tied to those signals, on the contrary, it was observed that education was a key proxy characteristic that the model observed, as was gender and income. The predictive model worked harder, favoring women and the 45-65 age group. There was a significant amount of bias against those with lower education as well. At this time there is no clarity on what drove this bias in the model. It could have been influenced by any number of data signals, for example, circumstances driven by signals consumed at impression time. Whatever the causes, the dynamic creative prediction model didn't work as hard to assemble the appropriate creative elements for lower education as it did for those with higher education levels. This could result in a reduced number of lower education audience members' conversions.

With multiple sub-groups surfacing as disadvantaged, or overtly advantaged groups, the metrics suggested that bias mitigation would foster broader, fairer exposure to messaging predicted to convert. In addition, the theory emerged that with appropriate mitigation strategies in place the models would work harder to predict more effective messaging and creative combinations for groups that were highly likely to convert. Mitigation itself, however, proved a little more complicated.

With the research focused on a post-processing environment, there was no live access to the predictive model and there were over 12,000 features to consider across the dataset. With this in mind the team employed a method that transformed the predicted probabilities from the machine learning model to mitigate bias based on the criteria available in the campaign data. However, inherent challenges with advertising campaign data still remain. Typically, the overall scale of an impressed audience is in the millions, but average click-through rates are only a smaller fraction of that population. The bias mitigation method works with heavily imbalanced data in this case, because there is less than a 0.01% chance of someone converting, especially within the identified groups. This means that the mitigation method doesn't have enough data to learn about converters to allow for mitigation. As a first step towards overcoming this imbalance, the number of non-converts in a group was reduced to nearly equal the converts, providing the bias mitigation method an opportunity to learn how to mitigate bias successfully. By adopting this common practice in cases of data with heavy imbalance, the bias mitigation algorithm was able to mitigate across multiple disadvantaged groups.

Another observation of note was that if mitigation for education and another proxy characteristic that had significant disparate impact such as political affiliation was in place, this sub-campaign could have benefited from a better-tuned predictive model. In addition, when those two biased signals were mitigated, a positive ripple effect for other biased signals was observed. As an example, a group within the proxy sub-group of religion improved by several disparate impact points as a result of mitigation between education and political affiliation. There is additional testing required to determine if these ripples can be expected in any scenario.

This bias mitigation methodology could also be utilized to positively influence a system where the model and training data are accessible. This can be achieved by pre-processing mitigation methods such as reweighing the instances in the training data. This might open possibilities for both similar and related campaigns to benefit from mitigation strategies as well as the opportunity for platform-based algorithms to provide categorical mitigation strategies, for example where a repeatable outcome will be common across industries.

"One of the most significant observations is: bias can exist in the data and algorithms that are employed for digital advertising, and that bias is not always immediately observable to the human eye."

Observations and findings

The work so far has revealed several learnings and opportunities about the presence of bias in advertising and led to additional paths for exploration. One of the most significant observations is: bias can exist in the data and algorithms that are employed for digital advertising, and that bias is not always immediately observable to the human eye.

Below are the most significant findings so far:

Bias exists.
Bias exists within both determined and proxy characteristics in advertising campaign data. The observation of proxy biases means that machine learning models may see more than just the targeting data provided. They have the ca pacity to react to other forces beyond the predefined inputs like a segment. Marketers need to fully understand the impact that bias could have on their consumers, their campaigns and their brands. The bias identification methods described above are one mechanism for identifying unintended bias, but there is more work to do to understand which external forces drive these inequalities.

Advantages and disadvantages are both important.
The team observed both overtly advantaged and disadvantaged groups, illustrating the need for mitigation methods to drive better performance and equal opportunity. Within the Ad Council sub-campaign, women and the age group 55-64 received significant advantages over the remainder of the audience, with that age group having a disparate impact ratio of 32.7 (way over 1) while the remaining groups fell at or below a mitigatable ratio (less than 1). Helping models balance these groups could allow better prediction of the most effective creative to drive conversion across a broader range of the target audience.

We should question typical campaign outcomes.

These tests have led to the question: “Are typical campaign reporting outcomes considering all of the possibilities?” Often the underlying machine enabled biases could cause malformed interpretation and further perpetuate strategies that induce unintended bias against the brand or consumer. The only way to truly understand these possibilities is to process campaign data differently and try to understand why a group might be highly performant, while others are not. The tools described above have high potential to aid in these types of explorations.

Mitigation is possible.

The team successfully tested mitigation of the impacts of unintended biases in post-campaign data. Test results suggest that mitigation strategies may positively impact campaign performance. As outlined above, these methodologies can be applied to a model before and during training as well as campaign run-time. While there is still work to be done, there is a strong possibility that mitigation strategies can help improve population fairness and increase campaign effectiveness.

Mitigation can have a waterfall effect.

Mitigation strategies might have a cascading effect, with the mitigation of bias for one group positively impacting another biased group. Initial tests in transformation illustrated that mitigation on education and political affiliation reduced the negative impacts on religion. This impact on additional groups also codifies those biases can be deeper, unobserved signals that intermingle within a model’s learning.

Humans are complex.

Bias scanning can provide a deeper understanding of audience composure and segment effectiveness. Humans are very complex and have many overlapping interests, emotions and states of being. The overlaps in interest-based segments and cohorts that humans exist within are, though blind to the human eye, appear to be caught by the machines. These overlaps can potentially cause an imbalance, and the bias scanning can help identify the areas of concern and uncover underlying connections to create better strategies.

The research continues...

While unintended bias has been identified and mitigated within digital advertising data in this test, there is still much work to be done and additional questions and theories to be tested. The Ad Council sub-campaign focused on tools like predictive dynamic creative optimization, bid optimization, and segmentation, but the digital advertising industry employs many different types of data and algorithms. This work is not meant to put any single brand or technology under the microscope as an audit of their choices.

In fact, the team largely believes that the majority of the biases that are being observed are not the result of malicious intent. Rather, they are the cognitive remnants of the strategies and decisions made by humans in the planning or development process, or possibly the impact of different platforms with various goals attempting to interact at massive speeds.

"The majority of the biases we'll encounter are not malicious in intent, but rather cognitive remnants from strategies and decisions made by humans in the planning or development process."

Making bias detection and mitigation even easier

The team is currently working with AIF360 to automate the complex techniques that have been applied to both bias detection and mitigation across datasets with tens of thousands of features. This Auto-AI service will allow automation of Multi-Dimensional Subset Scans to explore all features and feature overlaps for instances of systemic advantage or disadvantage and will automatically utilize fair score transformer test mitigation strategies seeking campaign data-wide fairness. This tooling could potentially help establish metrics that illustrate the impact of existing biases and how mitigation might optimize them.

Your campaign data is rich with opportunity

IBM Watson Advertising is continuing this research and is looking for partners that are willing to submit their campaign data, log files, and even models for exploration. A few examples of how brands can participate, along with the types of data needed, are outlined below:

1. A general campaign study would take a standard campaign and evaluate how the downstream algorithms are reacting to the underlying data. ***Requires:*** *campaign logs across ad-tech stack inclusive of impressions, conversions and identifiers*
2. A creative optimization study would explore, similar to the Ad Council campaign, what creative optimization algorithms are seeing in the data, and the choices they're making towards optimization. ***Requires:*** *campaign logs across ad-tech stack inclusive of impressions, conversions and identifiers, predictive propensity scores for optimization*
3. A First Party Audience study would explore whether the ways in which a brand employs and manipulates their FPD does not cause downstream impacts due to bias. ***Requires:*** *audience cleanroom data, methods/models for audience manipulation*

Action is necessary

As the advertising industry swiftly works to architect the ad-tech of tomorrow, every brand and agency should be taking steps to get educated on how bias works, how it can impact their teams, campaigns, and the technology they employ. Missing the opportunity to build trusted and equitable practices to demonstrate what brands can expect and consumers will experience could be detrimental to the industry and consumer trust.

IBM Watson Advertising will continue to lean into these tools and evaluate ways to reduce biases and increase the effectiveness of digital advertising technologies and practices. If you feel this work and its outcomes align with your organization's principles and approaches to consumer trust, please get in touch with our team to get involved in this effort.

To learn more about this research initiative and how you can participate, please visit: ibm.biz/bias-in-advertising

About IBM Watson Advertising

IBM Watson Advertising makes data actionable through a suite of privacy-forward, AI solutions that help brands make an impact on their business outcomes. Learn more at ibm.com/watson-advertising.

About IBM Research

IBM Research is a group of scientists, technologists, designers, and thinkers inventing what's next in computing. We're relentlessly curious about all the ways that computing can change the world. We're currently obsessed with advancing the state of the art in AI and hybrid cloud, and defining the future of quantum computing. We've been at the forefront of the computing revolution from the start: Our researchers have played a role in some of the most important advancements in technology, from the hard drive and the floppy disk to mainframes and the personal computer. Since our first lab opened in 1945, we've authored more than 110,000 research publications. Our researchers have won six Nobel Prizes, six Turing Awards, and IBM has been granted more than 150,000 patents.

—

Sources:
1. *https://www.statista.com/statistics/617136/digital-population-worldwide*
2. *https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai*
3. *https://www.mediapost.com/publications/article/367587/more-than-half-the-ad-industry-has-no-plans-to-dev.html*

Produced in the United States of America
January 2022