Advertising Playbook for AI Fairness 360

Watson Advertising

IBM

Contents

Introduction

The advertising industry constantly transforms and adapts to evolving consumer habits, emerging channels and tactics, and shifting societal requirements and expectations. Over the past few years, brands and marketers have been called on to live their values more than ever by employees, consumers, partners and shareholders, and specifically around diversity, equity, and inclusion. While these efforts to date have certainly helped transform DE&I practices, they have not impacted the fundamental underpinnings of advertising technology and data collection.

As we rely more and more on scalable, automated tools to help drive critical [advertising] decisions at scale, we must apply ethically grounded, trustworthy, and responsible AI at this epicenter of technology, brand, and human convergence. We should seek to employ AI governance in our strategies, campaigns, tools, and platforms that provide explainability, fairness, and robustness alongside transparency and privacy.

In the summer of 2021, IBM Watson Advertising initiated a research effort using tools developed by IBM Research to explore bias in advertising - to identify and mitigate bias within campaign data, algorithms, and outcomes. That research uncovered that discrimination could exist in the tools we employ.

This document and the accompanying tools, guides, and explainers intend to help the advertising industry put equal weight and importance on the machine learning tools and data techniques we employ, as we do on DE&I education and awareness. It's time to get into action. Through our collective efforts, we can help usher in a new era for the advertising industry balanced in privacy-forward, human-centered practices that strive to create fairer outcomes for brands and humans.

Roles in identifying and mitigating bias in advertising

The advertising ecosystem is a massive complexity of decisions, data, and outcomes co-mingled across many different organizations for one single campaign line item. The broad focuses and contributions that we all make, from brand leaders to strategists to developers, require that we all take notice, learn and take action on the potential of implicit bias in our campaigns, data and technology.

This document includes several sections, through which a broad understanding of topics and very technical approaches can be obtained.

Non-Technical Contributors

Every contributor to marketing and advertising should be aware of the negative impacts of unconscious bias on data and algorithms. From the Chief Marketing Officer to the performance analyst, they are increasingly responsible for deciding which ad tech and mar tech they'll deploy to understand where ads should be placed, predict which ads consumers are likely to engage with, which audiences are most likely to convert, and which publisher is attributed for a conversion.

A foundational understanding of how AI Fairness works can help drive better decisions across the campaign activation landscape and inform the vendor selection process with thoughtful consideration of how data will intermingle with functionality.

The first half of this document is for all contributors.

Technical Practitioners

The technical practitioner is an active participant in the design, development, architecture, and activation of a system that employs data and machine learning to drive an outcome. Often these teams are responsible for acting on the requirements and objectives of non-technical contributors. Similarly, they are also uniquely positioned participants who can employ fairness tools to construct informed outcomes.

It is essential for these practitioners to obtain a practical application of fairness strategies and tools for designing and developing new systems, and for the modernization of existing platforms. The concepts and approaches outlined in sections 2 and 3 below will be essential.

How to use this document

Advertising Playbook for AI Fairness 360 is broken into four parts, including a broad overview of fairness, technical explanations and functional guidance.

Part 1 - Guidance for the Future of Advertising: clarifies why fairness is crucial to advertising, how ethics by design can inform better outcomes, and key terms and considerations for any role across the advertising ecosystem.

Part 2 - Playbook for Practitioners, Designers, and Developers: introduces fairness by design concepts, the Advertising Toolkit for Fairness 360, and how to apply the tools in practice.

Part 3 - Practical Applications and Examples: provides an overview and key concepts in an application for a series of python notebooks showcasing critical steps and considerations. New examples will be added as additional studies are completed across the industry.

Part 4 - Moving beyond Bias Research: provides practical next steps, additional areas of education, and how participants can add content and findings to the Advertising Toolkit for AI Fairness 360.

How can you make a difference?

The most vital first step anyone across the advertising and marketing landscape can take is to get smarter about the issues. Diversity, equity, and inclusion training alongside a practical understanding of how biases can seep into our data and algorithms alone, can help drive industry-wide, inquiry-driven transformation. There are many resources available to introduce these concepts and industry organizations like the 4As and IAB are conducting agency and brand training programs. As an example, the 4As *Campaign Enlightenment* is an intensive four session workshop to help organizations set a DE&I vision and establish inclusive-based campaign development practices. Global agencies like WPP's Mindshare have developed specialty practices solely dedicated to creating more equitable and inclusive outcomes for people and growing brands through their Intentional Media Framework.

As a practitioner in the engineering, development and data science disciplines, one can explore the tools, apply them in practice, and consider opportunities to build them into processes and platforms–essentially changing the industry from the inside out.

Collectively as an industry (and individuals) we can engage in honest and transparent conversations with each other AND the consumers we seek to reach. In 2022, IBM Watson Advertising along with participation from advertising industry organizations, agencies and brands committed to an ongoing initiative to bring the industry together in action with the Advertising Fairness Pledge. Take the pledge, get educated, explore bias in your practices and be an advocate for your organization.

Part 1
Guidance for the Future of Advertising

IBM

Advertising is Essential

Advertising and marketing are essential to the economy and have the power to shape societal views and norms. Without it, businesses don't grow, jobs don't get created, and humans don't discover the things that make their lives interesting, different, and uniquely theirs. This influence, combined with the continued growth in digital channels and consumption models, has the potential to drive meaningful impact and potential discrimination at scale.

As the advertising industry re-evaluates and rebuilds the technology infrastructure around a foundation of trust with consumers, that trusted AI principles and transparency should be built-in.

The advertising industry has the power and responsibility to shape the future for a new generation who cares deeply about aligning with brands with a shared set of values. And this has to include identifying and mitigating bias.

DE&I Education is Leading the Transformation

Diversity, equity, and inclusion-focused education is the right place to start. There are many different paths to follow for individual and organizational transformation. Several programs focus specifically on the advertising industry, providing depth of coverage from strategy, user imagery, and addressable audiences, to how to hire a diverse workforce. This work is essential to the future of our industry.

If you're unfamiliar with DE&I training or practice development for the advertising industry, you might consider exploring resources being made available by the 4As, IAB, See Her, Female Quotient, and the ANA. Many of these resources are self-administered, while some programs provide a focused training effort, designed to transform an organization. You can explore some of these tools and resources in section 4.

Pushing the Right Buttons

There has been tremendous work done to educate and train the advertising industry and society to recognize the deep-seated prejudice stemming from inequality. But that same effort and rigor has not yet been applied to the advertising technology and data practices enabling the connection between brands and consumers. Yet, at the end of the day, we are all required to push a button somewhere, beginning a process that will take our strategies, our data, our good intentions and translate them into action. What happens beyond that action is just as important as all the work leading up to it.

This was the impetus for the Bias in Advertising Research work, to affirm what happens within the technology that we entrust with our data, our objectives, outcomes – as the final bridge between brands and their consumers.

The value in much of this work is focused on how important it is for brands to work on inclusionary strategies and technologies for the humans we seek to reach. But there is a great deal of value for the brand too. The work of understanding how the data we collect and the algorithms we employ might be impacted by bias has the potential to uncover new learnings and opportunities for how we strategize reaching them in the first place.

What do we mean by Bias?

To appropriately frame how we can change our ecosystem, we must define what we mean by bias and where things like unconscious or implicit bias come from. Additionally, as we seek to understand how these cognitive biases might end up "behind the button," we need to consider how bias looks through the lens of the machine or algorithm.

Bias is a well-traveled word. While the definition is roughly a "prejudice in favor of or against one thing, person or group with another," this idea can materialize in several forms depending on the situation. For example, data science and psychology biases take on tremendously different meanings. Generally, a bias in psychology is about how an individual's view of the world takes shape and the events that impact their experiences and opinions. Conversely, data science bias can take on many different forms that are not necessarily negative in impact. For example, time interval bias can be caused by purposefully selecting a range of times to impact the desired conclusion.

Bias is also human nature. It comes from our learned interpretation, desired outcomes, and proclivity for 'group think.' When we don't consider it within our actions, it finds its way into our outputs. When we let bias seep into our data and how that data is interpreted and manipulated—for instance, in an advertising campaign—the impact grows. The results of that campaign, and future strategies based on it, are all grounded in the same, potentially biased foundation. When we stand back and think about the collective impact over the broad and rich fifty-plus-year history of the advertising and marketing industries, there is reason to pause and self-reflect.

For our purposes of evaluating biases' impact on advertising technology, we are referring to bias as:

"prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair."

It can start with the human and is often unintentional.

"Advertising *is* inherently biased".
You might immediately react with feelings similar to the above sentiment. It references how our industry employs signals like demographics to help divide and organize humans into audiences. The conscious decision to separate individuals is intentionally biased, but these words are a bit misappropriated. The challenge is that bias does not begin in interpreting the data but in applying human-driven intelligence and strategy to the data.

While it is essential for advertising to have the mechanism for grouping people, for example, by their relevance to a specific product's benefits or features, this act must be measured and carefully employed. Even when an audience is intentionally selected, the possibility remains that systemic disadvantages could result due to unconscious bias.

Cognitive Biases

Cognitive biases are constraints in objective thought rendered from the proclivity for the human brain to perceive information via a filter of individual experience and preferences. The filtering process is heuristics, which is a coping mechanism enabling the brain to prioritize and process the expansive input it receives every second. These cognitive biases are often unconscious (or implicit) and applied in our day-to-day experience. Left unchecked, sometimes they can cause unintended outcomes and given the right mechanism, digital advertising, have the propensity to scale.

A few examples include:

Anchoring bias - An anchoring bias affects an individual or group's decisions when influenced by a particular reference point, expectation or anchor. Once the anchor is established, every consideration after that is influenced by that anchor, and might differ significantly from a decision not anchored in that belief.[1]

Bandwagon bias - Bandwagon bias is a kind of groupthink. It's a cognitive bias that encourages us to accept something because other people think it. As an example, bandwagon bias can make us believe something achievable is improbable because others have failed before us.[1]

Bias blind spot - A bias blind spot is recognizing the impact of biases on the judgment of others while failing to see the impact of biases on your own judgment. Blind spots tend to arise before, during, and after the development of a model or strategy—making them difficult to detect.[1]

Confirmation bias - Confirmation bias is an individual or group's tendency to interpret new data as confirmation of one's existing beliefs. This bias appears when we select the data or information that supports our opinions or beliefs, resulting in an echo chamber, while disregarding any additional data or views that offer contradictory outcomes.[1]

Halo effect - The halo effect occurs when positive opinions of one facet of a person, company, brand, or product bias the opinion of other aspects positively.[1]

Self-interest/Self-serving biases - When we observe positive events and successes through our character or actions but determine that negative results occurred because of external factors, this is called a self-interest bias.[1]

Ingroup/Outgroup bias - In-group bias is a social psychology pattern where a group supports its own members' beliefs, ideals, and actions over out-group members. In-group biases influence decisions and can be amplified within an algorithm to a broader out-group.[1]

Status Quo bias - When we refer to the current situation as the norm and deviation from that norm is considered a failure, we employ the emotion-based status quo bias.[1]

Not invented here bias - A not invented here (NIH) bias is typically a group-think effect that avoids information, data, or outcomes from outside sources. NIH can result from an unwillingness to accept competitive groups' data or outcomes because their results do not align with the group's requirements.[1]

[1] *If you'd like to learn more terms and concepts around cognitive biases and their potential impacts to the development of campaigns and technology, explore the IAB AI Working Group's "Understanding Bias in AI for Marketing, A Comprehensive Guide to Avoiding Negative Consequences with AI".*

Crafting a Trusted Ecosystem

The ability for artificial intelligence (AI) to perform important business tasks has grown by leaps and bounds in recent years. Its proliferation across the advertising industry has increased efficiency, and automated and scaled the effectiveness of everything from bid optimization to predictive audience building. As AI has progressed from a proof of concept to powering critical digital advertising workflows, it has become increasingly apparent that this general-purpose technology must be assessed in precise context for privacy, robustness, fairness, and explainability. These four assessments, along with transparency to stakeholders, constitute the five pillars of trustworthiness. If issues are discovered, they must be mitigated before serious harms occur.

But what are these pillars?

Privacy is the idea that personal sensitive information should neither be disclosed inadvertently nor when a system is breached by a malicious actor. Data privacy has been a growing area of focus and regulation for some time, and there are some nuances with AI in the mix. When models are being built off of a brand's first party data, it is important to consider the utilization of sensitive information.

Robustness is the ability of an AI system to remain accurate in different settings and conditions, including naturally occurring conditions and those set up by malicious actors to fool the AI. For example, a robust AI bid-optimization model will not completely fall apart at the outset of a major change in the world, such as a global pandemic.

Fairness ensures that an AI system does not yield systematic advantages to certain privileged groups and individuals (defined by characteristics such as gender and national origin) and systematic disadvantages to certain unprivileged groups and individuals. The predictive audience generation model should not systematically favor any group.

Explainability allows people to understand how (typically opaque) AI systems make their decisions. Brand owners, agency strategists, and media planners can all make sense of an explainable AI system, each toward its own goals.

Transparency is achieved when the various assessments along with their justifications are documented and presented to stakeholders. Factsheets containing assessments of accuracy, privacy, robustness, fairness, and explainability of the predictive audience generation model may be generated for model risk managers, regulators, and the general public.

What is Algorithmic Bias?

Algorithmic bias is the appearance of systematic and repeatable errors in systems that construct unfair outcomes, such as disadvantaging one group of users beneath the benefit provided to others.

The challenges with algorithmic bias are grounded in our societal dependence on algorithms to facilitate our constant engagement with and reliance on digital signals. Specific to the advertising industry is our near-infinite, ever-evolving stack of data chewing automation, prediction, and optimization toolings across every facet of the ecosystem. The potential impacts of biases within these systems are scaled by the sheer volume of impressions required to meet our needs.

Another consideration within the concept of algorithmic bias is the lack of widespread cultural, ethnic, gender, and disability diversity in developing algorithms that impact human consumptive experiences. While our DE&I approaches seek to enrich teams with more diverse viewpoints, the industry is already teetering on technology developed without these practices.

The possibility of algorithmic bias in advertising data and technology is real. We must seek education, research its presence, and develop practices to reduce its impacts.

Ethics in AI

While most of this document focuses on exploring bias and fairness in advertising and marketing, many of the concepts here touch the edges of an ethical code for what is appropriate in advertising. Though our intent is not to define a general ethical approach for the industry (ethics are something that each individual and organization must determine for themselves), some core concepts from existing ethical guidance apply.

Later in this document, we explore, at a general level, the importance of Fairness by Design and some ethical considerations. Below are a few critical considerations provided by thought leaders and AI ethics advocates. These principles are intended as guidelines for framing the ethical design of advertising systems built with AI.

The purpose of AI is to augment human intelligence.
Data and insights belong to their creator [meaning for example, that PII belongs to the individual]
Technology must be transparent and explainable.

IBM has been at the forefront of developing AI and AI policies for several decades. Ethics must be embedded in the design and development process from the beginning of AI creation for any system or feature. IBM AI Ethics leadership developed Everyday Ethics for Artificial Intelligence as a primer on how organizations, designers, and developers can adopt processes around ethical AI.

Agencies and organizations have been advocating for an ethical evolution to how data is employed. Mindshare and GroupM have been leading some of this work, positing that “just because we can, doesn’t mean we should” through their breakthrough Data Ethics Compass, launched with Unilever in 2021.

Fairness for AI

Algorithmic fairness is intertwined with social justice and cannot be reduced to technical-only conceptions. Fairness and justice are almost synonymous, and are political. There are several kinds of justice, including (1) distributive justice, (2) procedural justice, (3) restorative justice, and (4) retributive justice.

- Distributive justice is equality in what people receive—the outcomes.
- Procedural justice is sameness in the way it is decided what people receive.
- Restorative justice repairs a harm.
- Retributive justice seeks to punish wrongdoers.

All of the different forms of justice have important roles in society and sociotechnical systems. In the context of developing machine learning systems, teams need to focus on distributive justice. Since the objective functions of machine learning algorithms are formulated in terms of outcomes, distributive justice, which is also expressed in terms of outcomes, can be controlled.

The other kinds of justice are important in holistically tamping down racism, sexism, classism, ageism, ableism, and other unwanted discriminatory behaviors.

Basic Concepts in Fairness

In advertising and marketing we must utilize the tools available to find and deliver messages to the audience who will most benefit from the products or services that we are providing. We likely employ predictive and optimization-driven advertising technology tools to help us decide which individuals are most likely to convert or take another action. This sort of discrimination is acceptable and is the type of task machine learning systems are suited for. It becomes unacceptable and unfair when the predictive or optimization system gives a systematic advantage to certain privileged groups and individuals and a systematic disadvantage to certain unprivileged groups and individuals. Privileged groups and individuals are defined as those who have been more likely to receive a favorable label in a machine learning binary classification task. In digital advertising, an example favorable label could be 'most likely to convert, buy or subscribe.' These are assistive labels in that they provide a potential benefit to the consumer or the brand.

To help provide context, consider favorable labels in other domains like being approved for a loan, being hired, not being arrested, and granted bail. Within these examples, approval for a loan, being hired, and being granted bail are considered assistive favorable labels, while not being arrested is regarded as a non-punitive favorable label.

Protected Attributes
Privileged and unprivileged groups are delineated by protected attributes such as race, ethnicity, gender, religion, and age. There is no one universal set of protected attributes. They are determined by laws, regulations, or other policies governing a particular application domain in a particular jurisdiction.

In digital advertising we might arrive at these attributes through proxies based on interests or other classifications gathered through data-collection around consumer consumption and intent.
We will spend time on proxies later in this section.

Group and Individual Fairness
There are two main types of fairness to be concerned about: (1) group fairness and (2) individual fairness. Group fairness is the idea that the average classifier behavior should be the same across groups defined by protected attributes. Individual fairness is the idea that individuals similar in their features should receive similar model predictions. Individual fairness includes the special case of two individuals who are exactly the same in every respect except for the value of one protected attribute (this special case is known as counterfactual fairness). Given the prevailing and evolving regulations being rendered upon digital advertising and marketing, group fairness is the more important notion for you to consider, but you should not forget to consider individual fairness as well.

Representational Harms
Most of the use cases you will encounter will lend themselves to fairness in the context of direct allocation decisions, like selecting an individual as part of an audience or presenting an offer, but that is not the only possibility. There are also harms in representation or quality-of-service, such as bias in search results. As wider examples, image searches for professions might yield only white people, web search results for foreign-sounding names might be accompanied by advertisements for criminal defense attorneys, and natural language processing algorithms for language translation or query understanding might associate doctors with men and nurses with women automatically.

Representational harms are much more prevalent when dealing with non-tabular semi-structured data such as images, speech, and text because these modalities are typically first processed into compressed representations rather than directly input into classifiers. (In the case of deep neural networks, the early layers compute the representation.) Traditional statistical models that are not based on machine learning, and even deterministic models composed of a few hand-crafted rules, are typically most closely related to allocation and can be treated similarly to machine learning models in checking for fairness.

Some of the methods and techniques for allocative fairness can be used in representational fairness, but different techniques may be more appropriate. This is because it is often harder to clearly define the problem when dealing with representational harms – they may be quite varied and difficult to catalog.

Note that later in the document, we discuss representation biases resulting from the sampling process. Despite their similar terminology, representational harm and representation bias are different concepts.

What kinds of bias exist?

Unfairness in machine learning stems from unwanted bias. Where does bias come from? To answer this question, it is useful to imagine different spaces in which various abstract and concrete versions of the data exist: a constructed space, an observed space, a raw data space, and a prepared data space. The constructed space is an abstract, unobserved, theoretical space in which there are no biases. The constructed space is operationalized to the observed space through the measurement of features and labels. Data samples collected from a specific population in the observed space live in the raw data space. The raw data is processed to obtain the final prepared data to train and test machine learning models.

Social bias enters in the measurement process, representation bias enters the sampling process, and data preparation bias enters in the data preparation process, as shown in Figure 1. These biases build up and persist as the data moves along from left to right in the figure, all the way to the model training and predictions unless they are mitigated. Biases build up the same way in both the allocative and representational harm (remember this is a distinct concept from representation bias) cases.

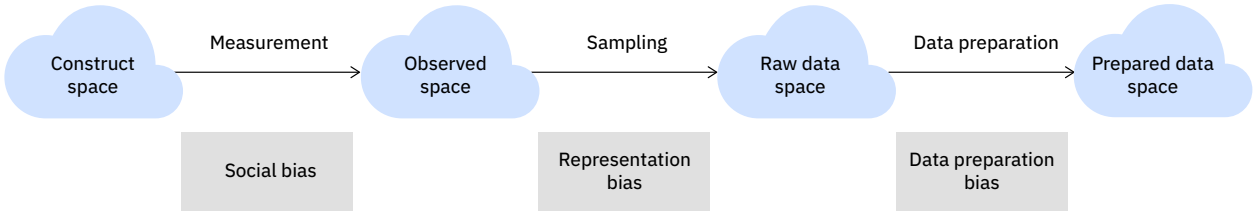


Figure 1: Different biases arising from different processes. (Varshney, Trustworthy Machine Learning, 2022)

Social Bias

Social bias may be in features or labels. Whether it is experts whose decision-making is being automated or it is crowd workers, people's judgment is involved in going from labels in the constructed space to labels in the observed space. These human judgments are subject to human cognitive biases which can lead to implicit social biases (associating stereotypes towards categories of people without conscious awareness) that yield systematic disadvantages to unprivileged individuals and groups. If decision-makers are prejudiced, they may also exert explicit social bias. For example, human data analysts may identify a potential conversion value for a man, but not a woman who is equally qualified, which shows up in training data labels. These biases are pernicious and reinforce deep-seated structural inequalities. Human cognitive biases in labeling can yield other sorts of systematic errors as well.

There can also be structural inequalities in features too. If an aptitude test asks questions that rely on specific cultural knowledge that not all test-takers have, then the feature will not, in fact, be a good representation of the test-taker's underlying aptitude. And most of the time, this tacit knowledge will favor privileged groups. Historical underinvestment and lack of opportunity among marginalized social groups also yield similar bias in features. Thus historical bias is a social bias.

Some more examples of social bias are the following. People who live in rural areas far away from doctors may be less likely to utilize the health care system for an equal level of infirmity compared to people in urban areas, so using health utilization as a feature for quality of health is filled with social bias. A feature that only counts income earned as salary rather than being inclusive of all kinds of income may have social bias against people who earn daily wages or own small businesses.

Representation Bias

Once operating in the observation space of features and labels, the data engineers on your team must actually acquire sample data points. Ideally, this sampling should be done in such a way that the acquired data set is representative of the underlying population. Often however, there is selection bias such that the probability distribution in the observed space does not match the distribution of the data points. A specific example of selection bias is unprivileged groups being either underrepresented or overrepresented in the dataset, which leads to machine learning models either ignoring their special characteristics to satisfy an average performance metric or focusing too much on them leading to systematic disadvantage.

Representation bias need not only be selection bias. Even if present, the characteristics of the features and labels that come from one subpopulation may be different than those from another. Representativeness is not only a question of the presence and absence of data points, but is a broader concept that includes, among others, systematic differences in data quality. For example, in a loan approval model that uses credit history as a feature, immigrants may have poorer quality data if their credit history from a different country is not accessible.

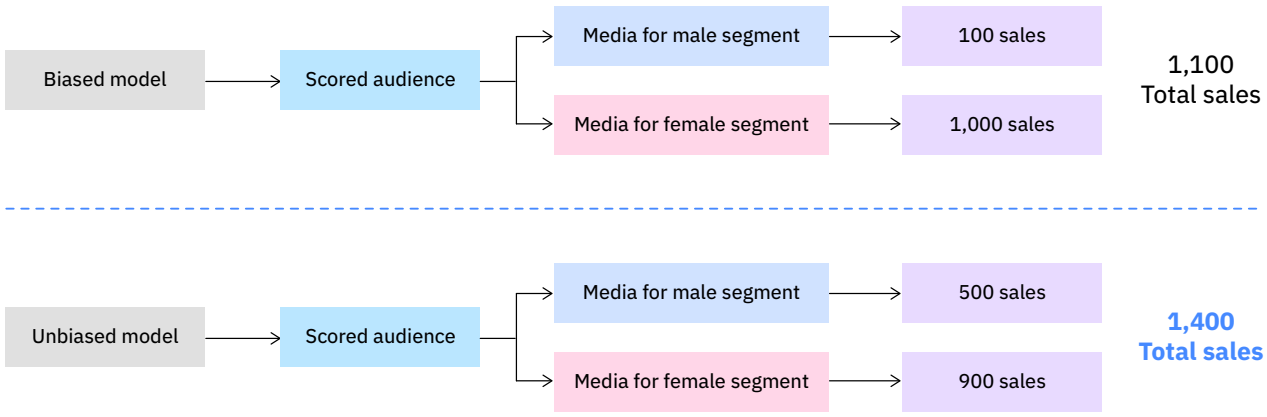
Data Preparation Bias

Many biases can be introduced in data preparation. For example, the data engineers on your team must do something to rows containing missing values. If they follow the common practice of dropping these rows and the missingness is correlated with a sensitive feature, like a debt feature being missing more often for certain religious groups, they have introduced a new bias. Other biases can enter in data preparation through other data cleaning, data enrichment, and data aggregation steps.

A sometimes overlooked bias is the use of proxies in the labels. For example, arrests are a problematic proxy for committing crimes. Innocent people are sometimes arrested and more arrests happen where there is more police presence (and police are deployed unevenly). Data preparation biases are often subtle and involve some choices made by the data engineer and data scientist, who are influenced by their own personal and social biases.

Okay, bias exists. What are the risks for advertisers?

The advertising industry prioritizes identifying signals that help to maximize outcomes. A key aspect of this process is leveraging historical data and context to help inform future strategies, but what the industry fails to recognize is the inherent bias that may exist in current approaches. Take for example, a situation where you have a biased model that helps identify prospective consumers to target. If the model relies on data that over-indexes on female shoppers, the prospective media targeting strategy will weigh more heavily for females. While female shoppers may result in more sales today, it inhibits our ability to identify and benefit from new sources of growth (e.g., sales) within the male population. By building an unbiased model, you create a more balanced representation of the prospective consumer pool.



NOTE: the above diagram is an illustrative approach, not an implemented solution or actualized results

This has implications for media investment strategy, audience and contextual targeting, creative production and personalization, partner selection, and measurement. Take the use case of bias in an audience strategy above, if you build a modeled audience based solely on optimizing for prediction power (e.g., propensity to drive sales) – not only could you limit your reach into new groups of consumers or behaviors that represent incremental sales in the short term, but you may hinder positive outcomes for other stakeholders (people, the industry, the planet) by perpetuating outdated, potentially harmful negative biases and in doing so you could actually experience tangible negative impact on your business long-term. This is central to what Mindshare calls 'Good Growth,' - growth that is enduring, diversified, and sustainable over time because it will align your business ambitions with the views and values of your consumers. The risk for marketers is the opportunity cost of not doing this – with potential implications for the company's financial performance, brand reputation, product strategy and innovation, and the possibility of affecting the communities they serve by feeding harmful biases into the zeitgeist through media. Having a tool that provides the ability to mitigate unintended negative consequences provides a much-needed level of accountability for marketers to society at large.

So, what should we do?

Education followed by explorative action is the best path toward a transformed digital advertising industry. While our collective goal should be to minimize the impacts of unconscious biases across all facets of the advertising and marketing ecosystem, realizing this goal will take time and considerable dedication. The foundation of those efforts is exposure to training and guidance towards new ways of thinking and the application of tools that are intended to support these actions. In Part 2 below, we will provide a deeper understanding for practitioners, guiding essential bias identification and mitigation concepts. However, the application of these tools will rely on our collective commitment to change.

Here are a few considerations for action:

1. Organizations should provide all teams with Diversity, Equity and Inclusion training. The outcomes of these training sessions will prepare teams for the right mindset to question existing practices and platforms and prepare them for future developments.

1. Explore cognitive biases and their cross-section with technological biases. Create a culture of questions around diverse points of view, helping to craft better strategies and develop better technology through an inclusive and understanding lens.

1. Ask teams to invest time and energy in learning tools, like the Fairness 360 toolkit discussed below to help them assess existing platforms and augment the design, development, and maintenance of future efforts.

A Phased Approach

There is no one size fits all solution. Depending on where individuals or organizations sit within the ecosystem, their approaches might be slightly different. The way to start the work of exploring bias in advertising technology is by asking questions. In identifying areas of potential concern, participants can craft problem statements and then orchestrate exploration into the problems by bringing together the appropriate teams and practitioners and applying the metrics and mitigation strategies outlined in Part 2.

The types of analysis that you might embark upon will vary depending on access to the advertising technology and data in question. Do you have access to the models? Do you own the data that is being collected? Organizations that understand the value of bias mitigation but are unsure of the potential impacts on how their technology operates may seek to explore previously trained data in a post-processing scenario. This could be achieved by examining the log-level data files from across a single campaign tactic or by evaluating the outcomes of a single tool within the advertising technology stack.

Later in Part 2, we provide in-depth coverage of where opportunities exist for bias mitigation in the machine learning pipeline; the application of these techniques will depend on the outcomes that the organization is seeking and should be discussed openly by stakeholders and practitioners. Brands, agencies, and advertising technology providers need to work together to achieve the best possible reduction in harmful bias across the ecosystem. Likely, brands will drive these efforts by looking to their direct and ecosystem partners to understand better how technology is processing consumer data. The relationship between brands, agencies, and ad-tech would position the brands as the ultimate stakeholders—their concerns and requirements will differ across verticals.

Our tactics have bias! Now what?

The work of researching bias in your advertising technology and data intends to discover when unconscious biases have impacted outcomes for brands and consumers. Given that the industry is operating on technology that is grounded in fifty-plus years of assumptions and tactics, it is likely that you will discover something.

Your findings may show a systemic disadvantage is rendered upon a marginalized or under-served group, perhaps surfacing concerns around gender, ethnicity, or income. You might also see that a disadvantaged group is not directly correlated with a protected feature like the above but is defined by proxy characteristics like homeownership, employment status, or interest in specific topics. The discovery of bias in the approaches you employ does not mean that your brand or partners are malicious in intent. Instead, it qualifies the necessity that we as an industry work harder to improve our processes and platforms with a thoughtful and inclusive approach.

You can use these findings to shift your strategies toward a more inclusive and impactful outcome while working with your industry partners to improve their tools towards results that will benefit your consumers' needs.

Part 2
Playbook for Practitioners, Designers & Developers

In Part 2, we'll dive into putting some of these concepts into action by exploring fairness and some of the metrics and mitigation algorithms available within the Advertising Toolkit for AI Fairness 360. While much of the content is formatted for the practitioner, the first few sections will greatly benefit all stakeholders. We encourage all readers to dig in.

IBM

While leaders and decision-makers across the advertising industry seek to drive meaningful change and progress in reducing unconscious and implicit biases in advertising technology, the responsibility of reducing bias in outcomes spans the entirety of the organization. Educational programs will provide awareness towards approaches that put reflection and questioning outcomes at the forefront of our processes. But in practice, the system designers, developers, engineers, and data scientists will enact both the work of assessment of current systems and the application of Fairness by Design to the systems of tomorrow.

This section is devoted to those practitioners who will need a fundamental understanding of how fairness metrics and mitigation algorithms can apply to advertising technology. Below many of these concepts are explored, providing essential information to help guide their utilization.

Fairness by Design

With the widespread implementation of AI and Machine Learning within the advertising industry, system designers and implementers must consider the possibility that human biases can be embedded in the systems we create. Whether driven by, for example, an intended outcome (confirmation bias), or the unconscious impacts of intersectionality, it is the role of the team responsible for the system to minimize algorithmic bias through continuous research and evaluation.

The team itself is a critical component of Fairness by Design. A diverse team can provide broader representation and variation in experience to help minimize bias. Leaders should build and empower teams with different characteristics and backgrounds, as a non-exhaustive list of examples: genders, ethnicities, ages, educational backgrounds, disabilities, familial status and cultural perspectives. These team members should be granted objection-free processes to drive empathetic and inclusive inquiry into designing a feature, function, or system.

A team can more effectively evaluate data with the appropriate diverse composure, identifying when research or collection methods may have been impacted by implicit racial, gender, ideological or other biases. Our teams can work together to design and develop without intentional biases and derive practices to avoid unintentional biases like stereotyping, confirmation bias, and sunk cost bias. They can conduct real-time analysis of algorithms to discover and bring light to both intentional and unintentional biases, explore their origin, and implement techniques to mitigate impacts.

Leaders should introduce approaches within their teams where each project begins with a series of questions helping frame fairness as a core KPI of the project.

- How can we identify or audit unintentional biases during the design and development of this feature?
- How can we ensure that our methods adapt to the change in our ongoing data collection?
- How can we derive a feedback mechanism to guide the correction of any unintentional biases in our design or decision-making process?

Through these shifts in approach, thoughtful questions, and multidisciplinary evaluations of the systems teams build, we can help guide the development of trustworthy AI across the advertising ecosystem. There are practical and conceptual approaches to utilizing AI Fairness 360 tools to aid teams in Fairness by Design in the following sections.

Leaders seeking additional guidance on Fairness by Design should explore the additional resources available in Part 4 - Moving Beyond Bias Research.

AI Fairness 360

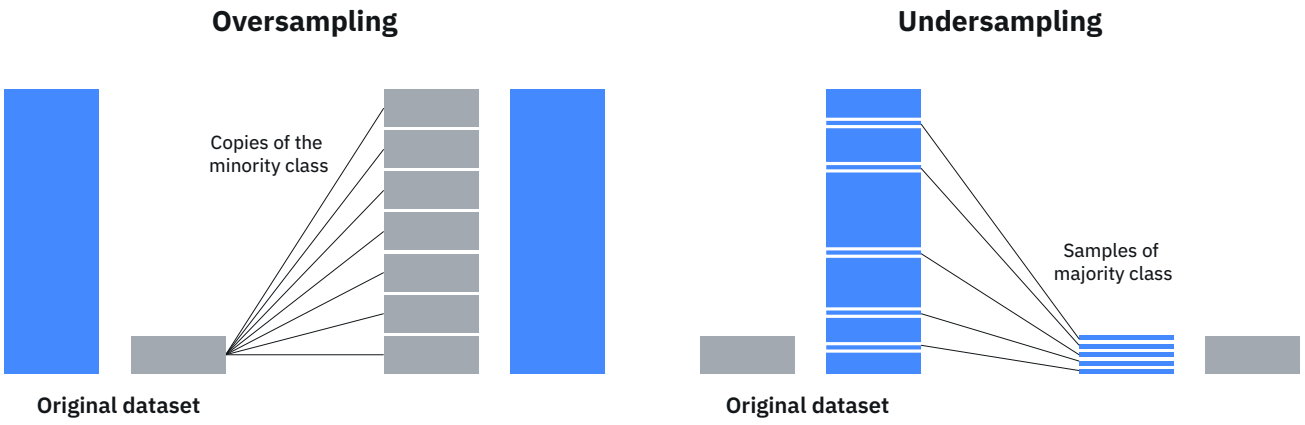
Machine learning models in advertising are increasingly used to inform fast-paced, high-value decisions about people, and while machine learning by its very nature is a form of statistical discrimination, that discrimination becomes objectional when it puts privileged groups at a systematic advantage and underprivileged groups at a systemic disadvantage.

AI Fairness 360 is an extensible open-source toolkit that can help brands, agencies, and ad-tech providers examine, report, and mitigate discrimination and bias in machine learning models used in any advertising technology that employs AI or Machine Learning. The toolkit includes 75 fairness metrics to help organizations identify the presence of bias and 13 state-of-the-art algorithms to mitigate discrimination and bias throughout the AI application lifecycle.

The mitigation algorithms within the toolkit include some of the top algorithms from experts within IBM Research, across the AI industry and throughout academia. Both the metrics and algorithms are outlined in tables 3 and 4 at the end of this section. Where available, the research papers are provided as links for further exploration.

Fairness Primer

Where does bias come from? Often we'll encounter the term "algorithmic bias" or "technological bias" and how we can unconsciously imbue a system with our cognitive biases. These cognitive biases, however, can also occur within the underlying data. Bias in data might happen due to over/under-sampling, label bias, user-generated bias, and poor, inefficient, or un-inclusive data collection practices. Our models are trained on data from our past human decisions, perhaps over many decades, and can reflect on societal or historical inequities.



As we seek to take the transition from concept to application, there are a few essential terms we will use that should be familiar:

Protected Attribute
An attribute that partitions an audience into groups whose outcomes should have parity (ex., Race, gender, income etc.).

Privileged Protected Attribute
A protected attribute value that indicates a group that has historically been at a systemic advantage (could be within in application scope or beyond)

Group Fairness
Groups are defined by protected attributes receiving similar treatments or outcomes.

Individual Fairness
Similar individuals receiving similar treatments or outcomes

Fairness Metric
A fairness metric measures unwanted bias in training data or models.

Favorable Label
A label whose value corresponds to an outcome that provides an advantage to the recipient

Below we'll continue into the concepts of fairness metrics assuming that you have already identified data sources, performed extraction, transformation, and loading into an environment, and are ready to explore.

Protected Attributes

One of the first things you should do with your data once it has been cleaned and prepared is to conduct exploratory data analysis. The standard basic approaches to conducting exploratory data analysis will sometimes reveal issues and unwanted biases to be aware of, but can miss subtle or diffuse patterns. As part of your exploratory data analysis, one approach you should use is specially designed to discover subtle anomalous patterns. It is known as multidimensional subset scanning.

You might not know which attributes to focus on as potentially protected attributes when you start. Although the decision to designate protected attributes is a policy decision and not a statistical decision, it is useful for data scientists to understand which subpopulations delineated by a convergence of different feature values have an exceptionally high or shallow value of the response variable compared to the population outside of the discovered subpopulation.

Multidimensional subset scanning, based on the linear time subset scan property, can efficiently find statistically significant anomalous subpopulations defined in terms of variables and their categorical or discretized values. It is efficient both statistically and computationally; it avoids the multiple testing problem and does not need to search through combinatorially many subgroups, which would be computationally prohibitive.

There are two variants to multidimensional subset scanning: auto stratification and bias scan. Auto stratification works on the prepared data before any classifier has been trained and will find the anomalous subgroup with either the provably highest average response variable value or the provably lowest average response variable value, depending on the setting. Bias Scan can be employed once the data scientists have created an initial classifier, only on the test set. Depending on the setting, it will find the subgroup with the provably worst error or the provably best accuracy.

The implementation in AI Fairness 360 also includes a penalty term that can be used to control the cardinality of the found subgroup (higher penalty means lower output length). Once one subgroup is found, it can be removed from the dataset, and the multidimensional subset scanning can be rerun to find the second most anomalous subgroup. The features forming the delineations of the subgroups should be surfaced to the problem owners as potential protected attributes or proxies.

In addition to identifying the subpopulation, multidimensional subset scanning also returns a score that is indicative of the statistical significance of the result. This score could then be converted into the p-value of the equivalent statistical hypothesis test using parametric bootstrapping.

Group Fairness Metrics

Now that you know the basic concepts of fairness and the unwanted biases that lead to unfairness in the sense of distributive justice and have conducted exploratory data analysis, it is important for you to check for fairness using quantitative metrics. Let's begin with group fairness in this section and come to individual fairness in the next section.

Group fairness is about comparing members of the privileged group and members of the unprivileged group on average.

Disparate Impact and Statistical Parity

One key concept in unwanted discrimination is disparate impact: privileged and unprivileged groups receiving different outcomes irrespective of the decision maker's intent and irrespective of the decision-making procedure. Statistical parity difference is a group fairness metric that quantifies disparate impact by computing the difference in selection rates of the favorable label $P(\hat{y}(X) = \text{fav})$ (rate of being exposed to advertising) between the privileged ($Z = \text{priv}$; e.g. male) and unprivileged groups ($Z = \text{unpr}$; e.g. female):

$$\text{statistical parity difference} = P(\hat{y}(X) = \text{fav} \mid Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Z = \text{priv}).$$

A value of 0 means that members of the unprivileged group (female) and the privileged group (male) are exposed to advertising messages at equal rates, which is considered a fair situation. A negative value of statistical parity difference indicates that the unprivileged group is at a disadvantage and a positive value indicates that the privileged group is at a disadvantage. An example calculation of statistical parity difference is shown in Figure 2.

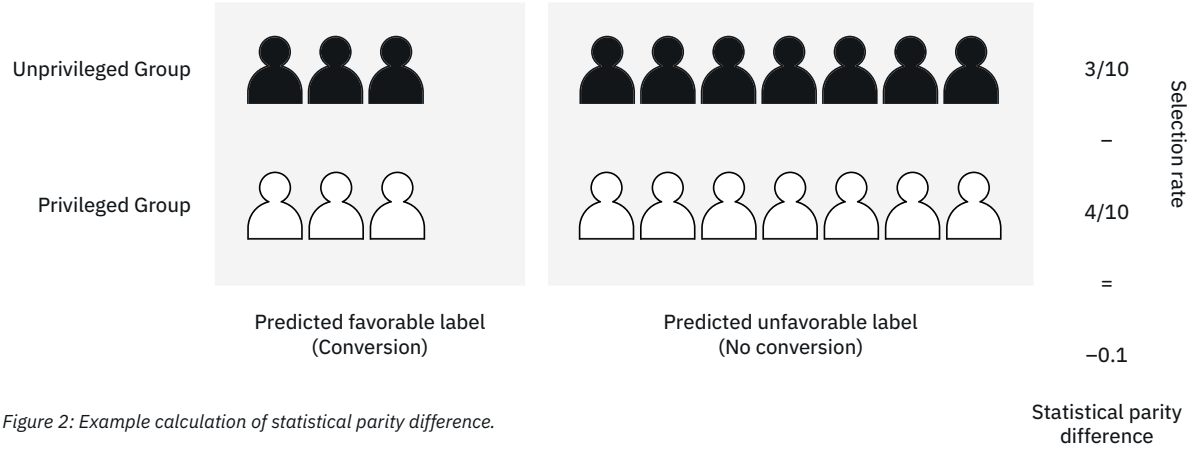


Figure 2: Example calculation of statistical parity difference.

Disparate impact can also be quantified as a ratio:

disparate impact ratio = $P(\hat{y}(X) = \text{fav} \mid Z = \text{unpr}) / P(\hat{y}(X) = \text{fav} \mid Z = \text{priv})$.

Here, a value of 1 indicates fairness, values less than 1 indicate disadvantage faced by the unprivileged group, and values greater than 1 indicate disadvantage faced by the privileged group. The disparate impact ratio is also sometimes known as the relative risk ratio or the adverse impact ratio. In some many domains, a value of the disparate impact ratio less than 0.8 is considered unfair and values greater than 0.8 are considered fair. This so-called four-fifths rule problem specification is asymmetric because it does not speak to disadvantages experienced by the privileged group. It can be symmetrized by considering disparate impact ratios between 0.8 and 1.25 to be fair.

Statistical parity difference and disparate impact ratio can be understood as measuring a form of independence between the prediction $\hat{y}(X)$ and the protected attribute Z. Besides statistical parity difference and disparate impact ratio, another way to quantify the independence between $\hat{y}(X)$ and Z is their mutual information.[2]

Both statistical parity difference and disparate impact ratio can also be defined on the training data instead of the model predictions by replacing $\hat{y}(X)$ with Y. They can then be measured and tested (1) on the dataset before model training, as a dataset fairness metric, as well as (2) on the learned classifier after model training as a classifier fairness metric, as shown in Figure 3.

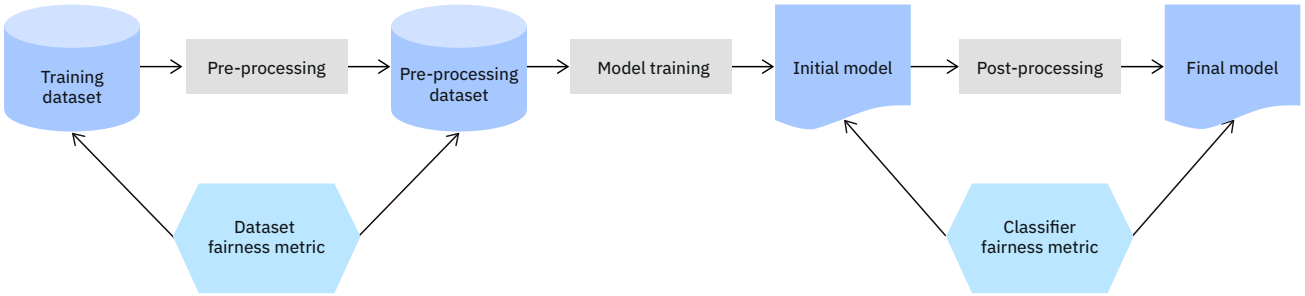


Figure 3: Points in the pipeline where dataset fairness metrics and classifier fairness metrics may be applied.

[2] *http://www.scholarpedia.org/article/Mutual_information*

While disparate impact ratio and statistical parity difference might provide useful insights about your models' performance, you might want to consider other group fairness metrics before attempting any mitigation.

Equality of Odds

A different group fairness metric is average odds difference, which is based on model performance metrics rather than simply the selection rate. (It can thus only be used as a classifier fairness metric, not a dataset fairness metric as shown in Figure 3.) The average odds difference involves the two metrics in the receiver operating characteristic (ROC): the true favorable label rate (true positive rate) and the false favorable label rate (false positive rate).

You take the difference of true favorable rates between the unprivileged and privileged groups and the difference of the false favorable rates between the unprivileged and privileged groups, and average them:

$$\text{average odds difference} = \frac{1}{2}\left(P(\hat{y}(X) = \text{fav} \mid Y = \text{fav}, Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{fav}, Z = \text{priv})\right) + \frac{1}{2}\left(P(\hat{y}(X) = \text{fav} \mid Y = \text{unf}, Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{unf}, Z = \text{priv})\right).$$

An example calculation of average odds difference is shown in Figure 4. The green circled plus signs indicate ground truth action, like clickthrough, view through or conversion.

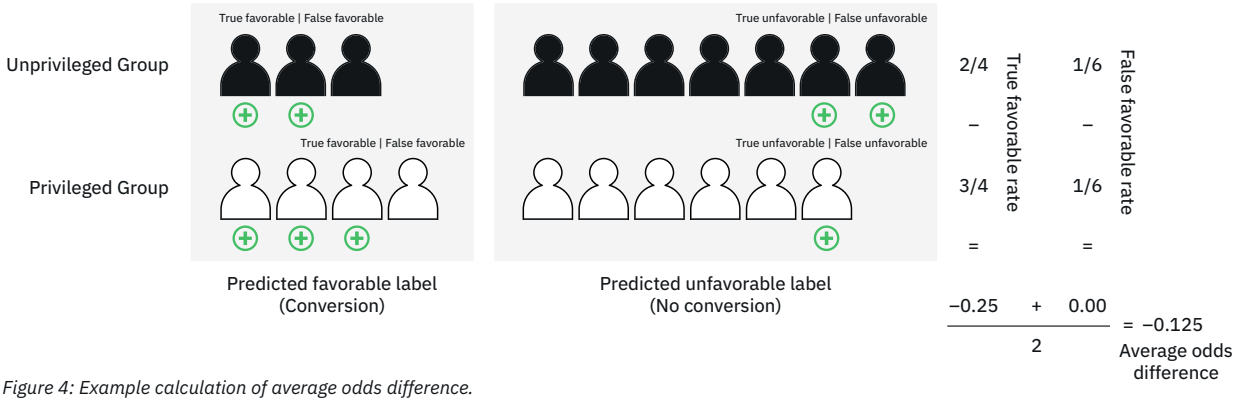


Figure 4: Example calculation of average odds difference.

In the average odds difference, the true favorable rate difference and the false favorable rate difference can cancel out and hide unfairness, so it is better to take the absolute value before averaging:

$$\text{average absolute odds difference} = \frac{1}{2}\left|P(\hat{y}(X) = \text{fav} \mid Y = \text{fav}, Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{fav}, Z = \text{priv})\right| + \frac{1}{2}\left|P(\hat{y}(X) = \text{fav} \mid Y = \text{unf}, Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{unf}, Z = \text{priv})\right|.$$

The average odds difference is a way to measure the separation of the prediction $\hat{y}(X)$ and the protected attribute Z by the true label Y in any of the three Bayesian networks[3] shown in Figure 5. A value of 0 average absolute odds difference indicates independence of $\hat{y}(X)$ and Z conditioned on Y. This is deemed a fair situation and termed equality of odds.

[3] *https://bayesian-intelligence.com/publications/bai/book/BAI_Chapter2.pdf*

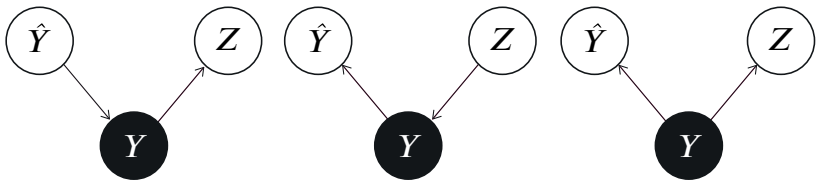


Figure 5: Illustration of the true label Y separating the prediction and the protected attribute in various Bayesian networks.

Choosing Between Statistical Parity and Average Odds Difference

What's the point of these two different group fairness metrics? They don't appear to be radically different. But they actually are radically different in an important conceptual way: either you believe there is social bias during measurement or not. These two worldviews have been named:

(1) "we're all equal" (the privileged group and unprivileged group have the same inherent distribution of addressability in the construct space, but there is bias during measurement that makes it appear this is not the case),

and

(2) "what you see is what you get" (there are inherent differences between the two groups in the construct space and this shows up in the observed space without a need for any bias during measurement).

Under the "we're all equal" worldview, there is already structural bias in the observed space, it does not really make sense to look at model accuracy rates computed in an already-biased space. Therefore, independence or disparate impact fairness definitions make sense and your problem specification should be based on them. However, if you believe that "what you see is what you get"—the observed space is a true representation of the inherent distributions of the groups and the only bias is sampling bias—then the accuracy-related equality of odds fairness metrics make sense. Your problem specification should be based on equality of odds instead.

Average Predictive Value Difference

And if it wasn't complicated enough, let's throw one more group fairness definition into the mix: calibration by group or sufficiency. For continuous score outputs of classifiers, the predicted score corresponds to the proportion of positive true labels in a calibrated classifier, or $P(Y = 1 \mid S = s) = s$. For fairness, you'd like the calibration to be true across the groups defined by protected attributes, so $P(Y = 1 \mid S = s, Z = z) = s$ for all groups z. If a classifier is calibrated by group, it is also sufficient, which means that Y and Z conditioned on S (or $\hat{y}(X)$) are independent. The Bayesian networks for sufficiency are shown in Figure 6. To allow for better comparison to Figure 5 (the Bayesian networks of separation), the predicted score is indicated by $\hat{Y}$ rather than S.

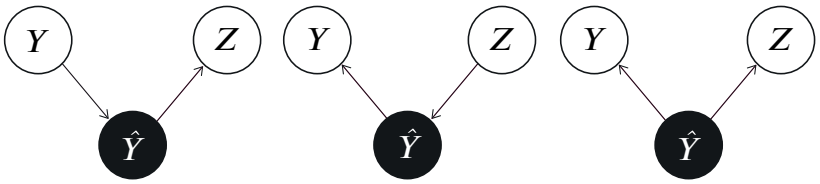


Figure 6: Illustration of the predicted label $\hat{Y}$ separating the true label and the protected attribute in various Bayesian networks, which is known as sufficiency.

Since sufficiency and separation are somewhat opposites of each other with Y and $\hat{Y}$ reversed, their quantifications are also opposites with Y and $\hat{Y}$ reversed. The positive predictive value is the reverse of the true positive rate: $P(Y = \text{fav} \mid \hat{y}(X) = \text{fav})$ and the false omission rate is the reverse of the false positive rate: $P(Y = \text{fav} \mid \hat{y}(X) = \text{unf})$. To quantify sufficiency unfairness, compute the average difference of the positive predictive value and false omission rate across the unprivileged and privileged groups:

average predictive value difference

$$= \frac{1}{2}\left(P(Y = \text{fav} \mid \hat{y}(X) = \text{fav}, Z = \text{unpr}) - P(Y = \text{fav} \mid \hat{y}(X) = \text{fav}, Z = \text{priv})\right)$$
$$+ \frac{1}{2}\left(P(Y = \text{fav} \mid \hat{y}(X) = \text{unf}, Z = \text{unpr}) - P(Y = \text{fav} \mid \hat{y}(X) = \text{unf}, Z = \text{priv})\right).$$

An example calculation for average predictive value difference is shown in Figure 7.
Again, the green circled plus signs indicate ground truth action.

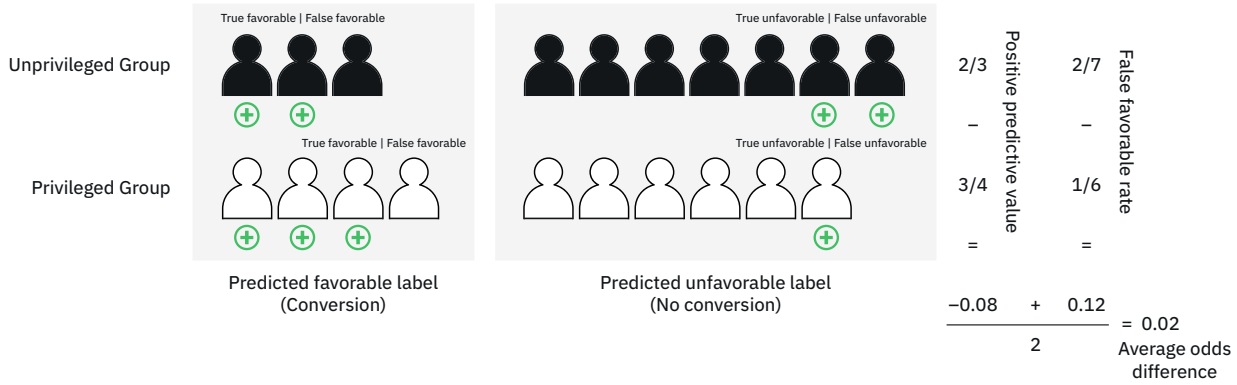


Figure 7: Example calculation of average predictive value difference.

The example illustrates a case in which the two halves of the metric cancel out because they have opposite sign, so a version with absolute values before averaging makes sense:

average absolute predictive value difference

$$= \frac{1}{2}\left|P(Y = \text{fav} \mid \hat{y}(X) = \text{fav}, Z = \text{unpr}) - P(Y = \text{fav} \mid \hat{y}(X) = \text{fav}, Z = \text{priv})\right|$$
$$+ \frac{1}{2}\left|P(Y = \text{fav} \mid \hat{y}(X) = \text{unf}, Z = \text{unpr}) - P(Y = \text{fav} \mid \hat{y}(X) = \text{unf}, Z = \text{priv})\right|.$$

Choosing Between Average Odds and Average Predictive Value Difference

What's the difference between separation and sufficiency? Which one makes more sense for your predictive conversion model? This is not a decision based on politics and worldviews like the decision between independence and separation. It is a decision based on what the favorable label grants the affected user: is it assistive or simply just non-punitive? Receiving an advertising message is assistive, but overexposing consumers to advertising and potentially causing negative outcomes, like over consumption, is non-punitive. In assistive cases like being exposed to a product's benefit, separation (equalized odds) is the preferred fairness metric because it relates to recall, which is of primary concern in these settings. In non-punitive cases like presenting offers that might drive consumers to buy additional products beyond their means, then sufficiency (calibration) is the preferred fairness metric because it relates to precision, which is of primary concern in these settings. (Precision is equivalent to positive predictive value, which is one of the two components of the average predictive value difference.)

Summary of Group Fairness Metrics

You can construct different group fairness metrics by computing differences or ratios of the various entries of the confusion matrix and other classifier performance metrics. These different permutations of group fairness metrics are implemented in the open-source AI Fairness 360 toolkit.

Despite this plethora of group fairness metrics, independence, separation, and sufficiency are the three main ones. They are summarized in Table 2, including when to choose each one.

Table 2: The three main group fairness metrics and when they should be selected.

Type	Statistical Relationship	Fairness Metric	Can Be A Dataset Metric?	Social Bias in Measurement	Favorable Label
independence	$\hat{Y} \perp\!\!\!\perp Z$	statistical parity difference	yes	yes	assistive or non-punitive
separation	$\hat{Y} \perp\!\!\!\perp Z \mid Y$	average odds difference	no	no	assistive
sufficiency (calibration)	$Y \perp\!\!\!\perp Z \mid \hat{Y}$	average predictive value difference	no	no	non-punitive

Individual Fairness Metrics

An essential concept in fairness is intersectionality. Things might look fair when looking at different protected attributes separately, but when you define unprivileged groups as the intersection of multiple protected characteristics, such as black women, group fairness metrics show unfairness. You can imagine making smaller and smaller groups by including more and more attributes, all the way to a logical end of groups that are just individuals that share all of their feature values.

At this extreme, the group fairness metrics described in the previous section are no longer meaningful, and a different notion of sameness is needed. That notion is individual fairness or consistency: all individuals with the same feature values should receive the same predicted label, and individuals with similar features should receive similar predicted labels.

The consistency metric is quantified as follows:

$$\text{consistency} = 1 - \frac{1}{n}\sum_{j=1}^{n}\left|\hat{y}_j - \frac{1}{k}\sum_{j\prime\in\mathcal{N}_k(x_j)}\hat{y}_{j\prime}\right|.$$

For each of the n potential customers, the prediction $\hat{y}j$ is compared to the average prediction of the k nearest neighbors. When the predicted labels of all of the k nearest neighbors match the predicted label of the person themselves, you get 0. If all of the nearest neighbor predicted labels are different from the predicted label of the person, the absolute value is 1. These people are ones who have poor individual fairness. Overall, because of the 'one minus' at the beginning of the equation, the consistency metric is 1 if all similar points have similar labels and less than 1 if similar points have different labels.

The biggest question in individual fairness is deciding the distance metric by which the nearest neighbors are determined. Which kind of distance makes sense? Should all features be used in the distance computation? Should protected attributes be excluded? Should some feature dimensions be corrected for in the distance computation? These choices are where worldviews come into play.

Typically, protected attributes are excluded, but they don't have to be. If you believe there is no bias during measurement (the "what you see is what you get" worldview), then you should simply use the features as is. In contrast, suppose you believe that there are structural social biases in measurement (the "we're all equal" worldview). In that case, you should attempt to undo those biases by correcting the features as they're fed into a distance computation. As an example, in advertising auctions where if you believe that pricing is fair, if the algorithm itself sees differences between women and men and their depth of experience, then the distance metric could help adjust for offsets, correcting the imbalance and evenly distributing their weights.

Counterfactual Fairness

One special case of individual fairness is when two consumers have exactly the same feature values and only differ in one protected attribute. Think of two consumers, one black and one white who have an identical brand engagement history. The situation is deemed fair if both receive the same predicted label—either both are given an offer or both are given the standard pricing—and unfair otherwise.

Now take this special case a step further. As a thought experiment, imagine an intervention that changes the protected attribute of a consumer from black to white or vice versa. If the predicted label remains the same for all members, the classifier is counterfactually fair. (Of course, actually intervening to change a consumer's protected attribute is not possible, but this is just a thought experiment.) Counterfactual fairness can be tested using treatment effect estimation methods.[4] Specifically, if there is no statistically significant effect of the protected attribute on the predicted label, while controlling for all possible confounding variables, then the situation is counterfactually fair.

[4] *https://ci360.mybluemix.net/*

Group and Individual Fairness Together

If you don't want to decide between group and individual fairness metrics, do you have any other options? Yes you do. You can use the Theil index, which was originally developed to measure the distribution of wealth in a society. It naturally combines both individual and group fairness considerations. A value of 1 indicates a totally unfair society where one person holds all the wealth and a value of 0 indicates an egalitarian society where all people have the same amount of wealth:

$$\text{Theil index } = \frac{1}{n}\sum_{j=1}^{n}\frac{b_j}{\overline{b}}\log\frac{b_j}{\overline{b}}.$$

The equation averages the benefit divided by the mean benefit $\overline{b}$, multiplied by its natural log, across all people.

That's all well and good, but benefit to who and under which worldview? The group that proposed using the Theil index in algorithmic fairness suggested that b_j be 2 for false favorable labels (false positives), 1 for true favorable labels (true positives), 1 for true unfavorable labels (true negatives), and 0 for false unfavorable labels (false negatives). This recommendation is seemingly consistent with the "what you see is what you get" worldview because it is examining model performance, assumes the costs of false positives and false negatives are the same, and takes the perspective of affected consumers who might meet the criteria to see an ad, even if they'll not convert. But this choice of b_j values is usually not appropriate from a societal or business perspective because it isn't sensible to highly reward false positives in which the advertiser loses money. More appropriate benefit functions for the audience optimization problem may be b_j that are (1) 1 for true favorable and true unfavorable labels and 0 for false favorable and false unfavorable labels ("what you see is what you get" while balancing societal needs), or (2) 1 for true favorable and false favorable labels and 0 for true unfavorable and false unfavorable labels ("we're all equal").[5] Case 1 corresponds to "what you see is what you get" because there is a benefit to being correct only when the prediction is correct in comparison to the observed space label.

[5] *Note that the scikit-learn-compatible API for AI Fairness 360 takes b_j as a parameter: https://aif360.readthedocs.io/en/latest/modules/generated/aif360.sklearn.metrics.theil_index.html, but the non-scikit-learn-compatible API does not, and defaults to the values suggested by the original group that proposed this metric (b_j = 2 for false favorable labels, b_j = 1 for true favorable labels, b_j = 1 for true unfavorable labels, and b_j = 0 for false unfavorable labels), which may not be the best choice: https://aif360.readthedocs.io/en/latest/modules/generated/aif360.metrics.ClassificationMetric.html#aif360.metrics.ClassificationMetric.theil_index*

Case 2 corresponds to “we’re all equal” because there is a benefit of the applicant receiving the favorable label (similar to the selection rate appearing in the disparate impact ratio).

An alternative metric that considers the path from group fairness to individual fairness is known as rich subgroup fairness. It starts with a group fairness metric such as statistical parity difference or average odds difference. Then just like multidimensional subset scanning implicitly considers all of the possible combinatorially many subgroups, the rich subgroup fairness metric does the same. It computes the average group fairness metric across all the possible subgroups, thereby ensuring that no one subgroup is treated unfairly. For this metric, you do not have to select protected attributes in advance. It also, in some sense, avoids issues with characterizing fairness when there are sampling problems, because it will consider all subpopulations, no matter how well or poorly they are sampled.

Individual Fairness Metrics Summary

Individual fairness consistency, Theil index, and rich subgroup fairness are excellent ways to capture various nuances of fairness in different contexts. Just like group fairness metrics, they require you to clarify your worldview and aim for the same goals in a bottom-up way. As the advertising industry considers privacy and fairness forward one-to-one marketing tactics, using group fairness metrics in your problem specification and modeling can be beneficial. Counterfactual or causal fairness is a strong requirement from the perspective of the philosophy and science of law, but the regulations are only just catching up. So you might need to utilize causal fairness in problem specifications in the future, but not just yet.

Bias Mitigation

Given the quantitative definitions of fairness and unfairness we’ve worked through, we know that bias results from some sort of statistical dependence between protected attributes like race, gender, or age and true or predicted labels like income, employment status, or homeownership. Bias mitigation should involve introducing statistical independence between protected attributes and labels. That sounds easy enough, so what’s the challenge?

Bias mitigation methods must be a little more clever than simply dropping protected attributes.

Proxies

What makes bias mitigation difficult is that other regular predictive features X have statistical dependencies with the protected attributes and the labels (a node for X was omitted from Figure 5 and Figure 6, but it really should have been there). The regular features can reconstruct the information contained in the protected attributes and introduce dependencies, even if you do the most obvious thing of dropping the protected attributes from the data. For example, race can be associated both with certain consumption behaviors (which may be a regular feature) and with historical brand interactions. These other regular features could then be proxies for the protected attributes.

One may desire to list all of the proxies for a protected attribute, like the neighborhood they live in, the types of products they purchase, who they engage with on social media, and so on. However this is a fruitless exercise because there is typically at least some weak dependence between most regular predictive features and the protected attributes. It is also not that helpful to list proxy variables because bias mitigation algorithms operate in ways that either figure them out on their own or do not need to know the proxy variables to achieve their bias mitigation goals.

Figure 8 shows three different points of intervention for bias mitigation: (1) pre-processing which alters the statistics of the training data, (2) in-processing which adds extra constraints or regularization terms to the learning process, and (3) post-processing which alters the output predictions to make them more fair. Pre-processing can only be done when you have the ability to touch and modify the training data. Since in-processing requires you to mess with the learning algorithm, it is the most involved and least flexible. Post-processing is almost always possible and the easiest to pull off. However, the earlier in the pipeline you are, the more effective you can be.

Figure 8: The three categories of bias mitigation algorithms apply at different parts of the machine learning pipeline.

There are several specific methods within each of the three categories of bias mitigation techniques (pre-processing, in-processing, post-processing). Just like for accuracy, no one best algorithm outperforms all other algorithms on all datasets and fairness metrics. In choosing a bias mitigation algorithm, in addition to knowing which part of the lifecycle you can touch, you have to (1) consider your worldview and (2) understand whether protected attributes are allowed as features and will be available in the deployment data. The deployment data is unlabeled data that you do not have access to when you are training and testing the model. It is the data that you encounter in production.

Pre-Processing

At the pre-processing stage of the modeling pipeline, you don't have the trained model yet. So pre-processing methods cannot explicitly include fairness metrics that involve model predictions. Therefore, most pre-processing methods are focused on the "we're all equal" worldview, but not exclusively so. There are several ways for pre-processing a training data set:

1. augmenting the dataset with additional data points

2. applying instance weights to the data points, and

3. altering the labels.

One of the simplest algorithms for pre-processing the training dataset is to append additional rows of made-up consumers that do not really exist. These imaginary individuals are constructed by taking existing member rows and flipping their protected attribute values (like counterfactual fairness; e.g. male to female and vice versa). The augmented rows are added sequentially based on a distance metric so that 'realistic' data points close to modes of the underlying dataset are added first. This ordering maintains the fidelity of the data distribution for the learning task. A plain uncorrected distance metric takes the "what you see is what you get" worldview and only overcomes sampling bias, not measurement bias. A corrected distance metric takes the "we're all equal" worldview and can overcome both measurement and sampling bias. This data augmentation approach needs to have protected attributes as features of the model and they must be available in deployment data.

Another way to pre-process the training data set is through sample weights. The reweighing method is geared toward improving statistical parity ("we're all equal" worldview), which can be assessed before the machine learning model is trained and is a dataset fairness metric. The goal of independence between the label and protected attribute corresponds to their joint probability being the product of their marginal probabilities. This product probability appears in the numerator and the actual observed joint probability appears in the denominator of the weight:

$$w_j = \frac{p_Y(y_j)p_Z(z_j)}{p_{Y,Z}(y_j, z_j)}.$$

Protected attributes are required in the training data to learn the model, but they don't have to be part of the model or the deployment data.

Where data augmentation and reweighting do not change the training data you have from historical consumer interactions, other methods do. One simple method, only for statistical parity and the "we're all equal" worldview, known as massaging, flips unfavorable labels of unprivileged group members to favorable labels and favorable labels of privileged group members to unfavorable labels. The chosen data points are those closest to the decision boundary that have low confidence. Massaging does not need to have protected attributes in the deployment data.

Data augmentation, reweighing, and massaging all have their own domains of competence. Some perform better than others on different fairness metrics and dataset characteristics. You'll have to try different ones to see what happens on your data.

In-Processing

In-processing bias mitigation algorithms are straightforward to state, but often more difficult to actually optimize. The statement is as follows: take an existing risk minimization supervised learning algorithm, such as:

$$\hat{y}(\cdot) = arg\ \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} L\left(y_j, f\left(x_j\right)\right) + \lambda J(f)$$

and regularize or constrain it using a fairness metric. In the case of regularization, $J(f)$ would be a fairness metric. The algorithm can be logistic regression and the regularizer can be statistical parity difference, in which case you have the prejudice remover. More recent fair learning algorithms, such as the meta-fair classifier, are broader and allow for any standard risk minimization algorithm along with a broad set of group fairness metrics as constraints that cover the different types of fairness.

Once trained, the resulting models can be used on new unseen audience members. These in-processing algorithms do not require the deployment data to contain the protected attribute. The trick with all of them is structuring the bias mitigating regularization term or constraint so that the objective function can tractably be minimized through an optimization algorithm.

Post-Processing

If you're in the situation that the model has already been trained and you cannot change it or touch the training data (for example if you are purchasing a pre-trained model from a vendor to include in your pipeline), then the only option you have is to mitigate unwanted biases using post-processing. You can only alter the output predictions $\hat{Y}$ to meet the group fairness metrics you desire based on your worldview (i.e. flipping the predicted labels from receiving the ad to not receiving the ad and vice versa). If you have some validation data with labels, you can post- process with the "what you see is what you get" worldview. You can always post-process with the "we're all equal" worldview, with or without validation data.

Since group fairness metrics are computed on average, flipping any random audience member's label within a group is the same as flipping any other random member's (this is what is done in equalized odds post-processing and calibrated equalized odds post-processing). A random selection of people, however, seems to be procedurally unfair. To overcome this issue, similar to massaging, you can prioritize flipping the labels of members whose data points are near the decision boundary and are thus low confidence samples (this is the reject option algorithm). All of these approaches require the protected attribute in the deployment data.

Individual Fairness Bias Mitigation

With the exception of the individual distortion constraint in the optimized pre-processing algorithm, the bias mitigation algorithms discussed thus far are focused exclusively on group fairness. There are, however, individual fairness-oriented bias mitigation algorithms as well. For example, an individual fairness post-processing algorithm constructs a graph of data points as nodes and edges between nodes within the neighborhood based on a defined distance metric. It then smooths out the discrepancies in the predictions of the graph using a mathematical object known as the graph Laplacian.

Also, the rich subgroup fairness metric, which is in-between group and individual fairness has a related in-processing algorithm known as the GerryFair classifier which adds a rich subgroup regularization term to an existing empirical risk minimization formulation.

Bias Mitigation Summary

All of the different bias mitigation algorithms are options as you're deciding what to finally do in the modeling pipeline. The things you have to think about are:

1. where in the pipeline can you make alterations (this will determine the category pre-, in-, or post-processing)

2. which worldview you've decided with the problem owner (this will disallow some algorithms that don't work for the worldview you've decided)

3. whether the deployment data contains the protected attributes (if not, this will disallow some algorithms that require them).

These different decision points are summarized in Table 3 for the bias mitigation algorithms available in the open-source AI Fairness 360. After that, you can just go with the algorithm that gives you the best quantitative results.

Fairness–Accuracy Tradeoff

Once you're building your models, deciding on the specific fairness metric you'll use, checking them for those fairness metrics, and mitigating unwanted bias, you might ask, shouldn't I consider a tradeoff of fairness and accuracy? Before getting there, it is important to note one important point. Even though it is what everyone does, measuring classification accuracy on the test set of data from the prepared data space, which already contains social bias, representation bias, and data preparation bias, is not exactly the right thing to do without mitigating those biases first. You should measure accuracy after bias mitigation (to be done in prepared space, since in construct space, being the real space, there is no unfairness). This means that a tradeoff between fairness and accuracy makes sense in prepared data space and not in construct space. For example, if in the construct space, men and women are equally targeting-worthy, but in the observed space they are not (due to social biases in the labels introduced by humans), your accuracy measurement will be flawed because it will count a predicted conversion for a women as incorrect when her construct label was convert and her observed space label was non-convert.

Practically speaking, you can approximate a construct space test set by using the data augmentation pre-processing method on the test set. But since it is not a widespread approach to do so, you should go ahead and look at the tradeoff between plain accuracy measured in the test set of data in the prepared data space and a chosen fairness metric, as starting point, before applying bias mitigation; then should also judge different bias mitigation algorithms empirically by their fairness–accuracy tradeoff curves, which are created by varying their hyperparameters. One example for one dataset is shown in Figure 9.

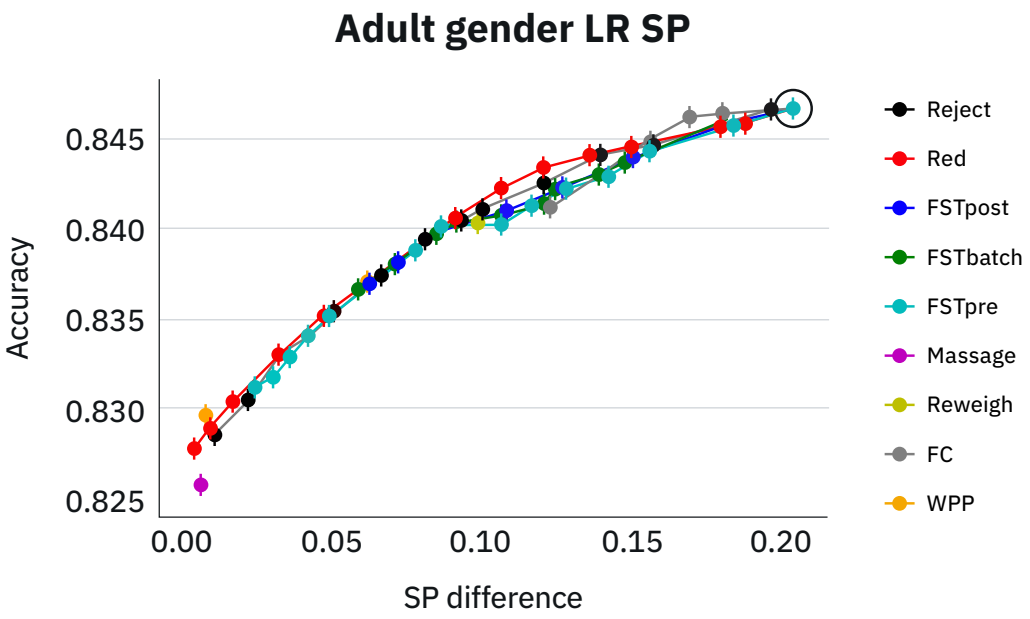


Figure 9: An example tradeoff curve between accuracy and statistical parity difference for several different bias mitigation algorithms.

Feasible Machine Learning Models

The relationship between fairness and accuracy is a key consideration in determining the acceptable quantitative ranges for fairness metrics that you set for your system. Without a sensible understanding of what is possible, policymakers may ask for a system with an accuracy of 100% and an average absolute odds difference of 0, which is only wishful thinking. Moreover, it is not only the relationship between fairness and accuracy that matters, but also metrics from the other pillars of trustworthiness mentioned in part one of this document, such as privacy, robustness, and explainability. All of these dimensions have interrelationships.

As schematically illustrated in Figure 10 using a few metrics such as faithfulness (an explainability metric), empirical robustness (a metric for adversarial robustness), and Brier score (a metric for predictive performance that also indicates calibration) along with accuracy and disparate impact ratio: there are tradeoffs among some dimensions and no tradeoffs among others. In the schematic, the set of metric values for one model maps to a single point on the plot. You should only dare to ask policymakers for acceptable ranges of values within the feasible region.

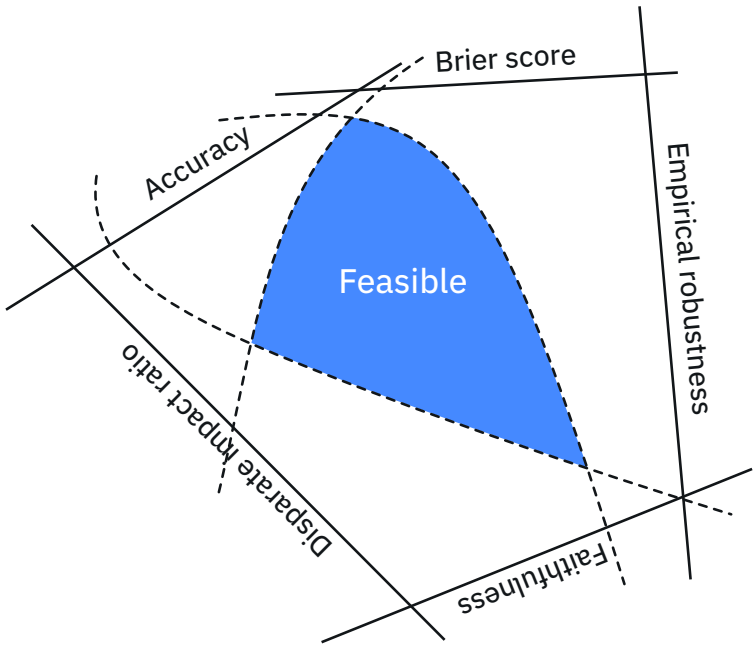


Figure 10: A schematic diagram showing the feasible region of machine learning models within which acceptable ranges should lie.

Elicitation

The feasible set is a good starting point, but there is still the question of deciding on the actual preferred acceptable ranges of the metrics. Two approaches may help. First, if you train many models for the same or similar prediction task and compute their performance metrics, you will get an empirical characterization of the interrelationships among the metrics. You can then better understand your choice of metric values based on their joint distribution in this set. The joint distribution can be visualized using simple bar graphs (shown in Figure 11) , a parallel coordinate plot (shown in Figure 12),[6] a parallel coordinate density plot, or a radar chart (shown in Figure 13).

[6] *https://github.com/IBM/conditional-parallel-coordinates*

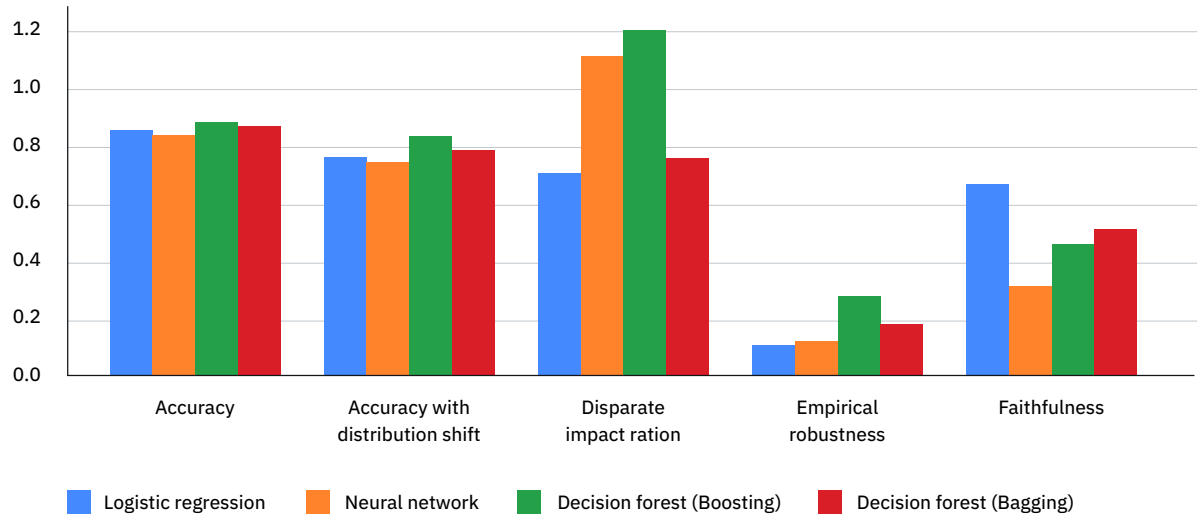


Figure 11: An example set of bar graphs showing performance metrics for four different models that let you understand the interrelationships among the metrics.

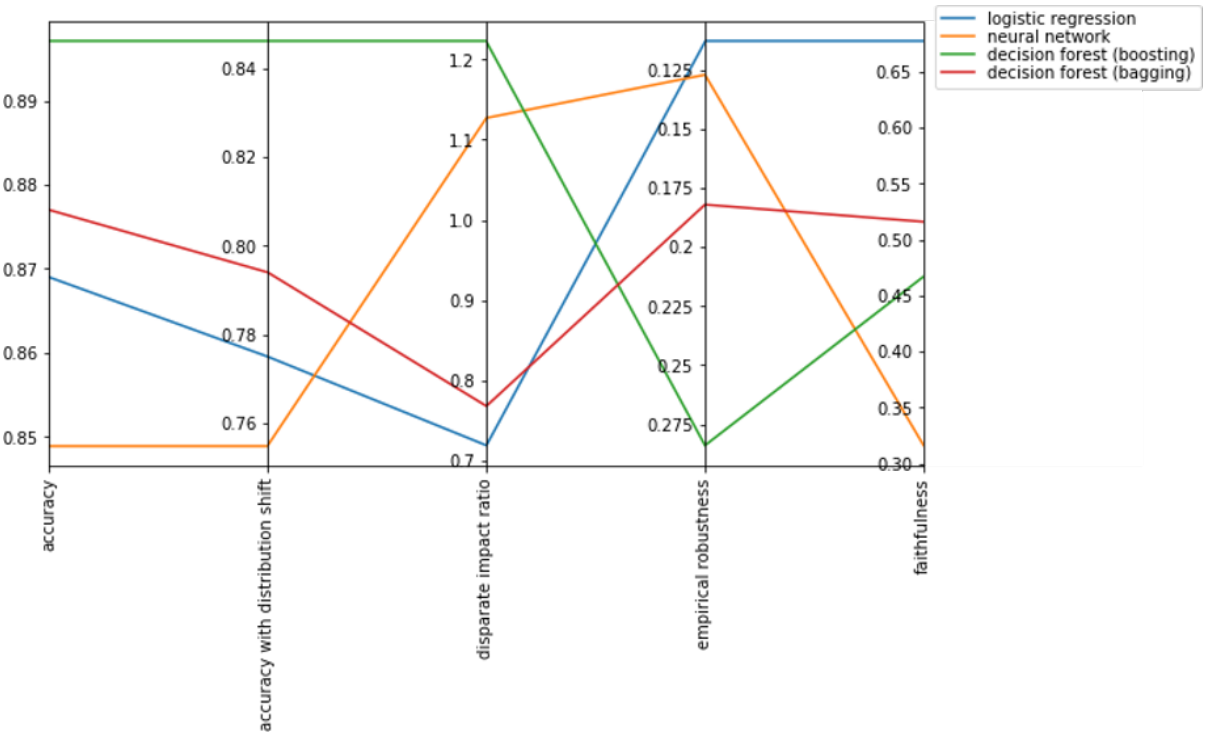


Figure 12: An example parallel coordinate plot showing performance metrics for four different models that let you understand the interrelationships among the metrics.

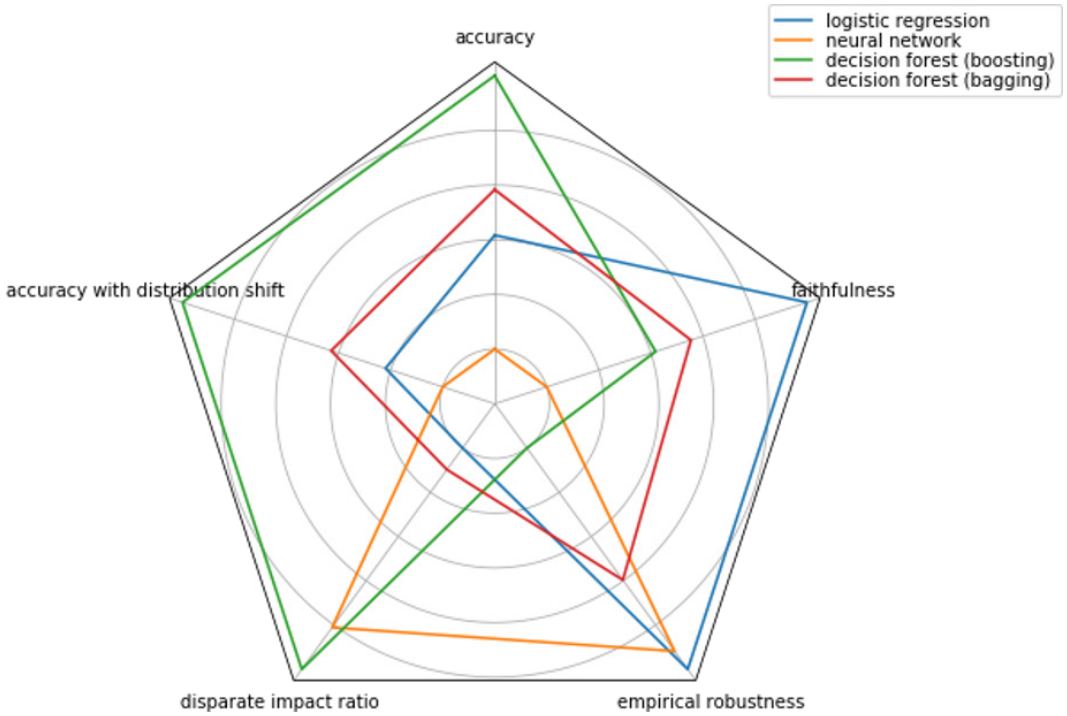


Figure 13: An example radar chart showing performance metrics for four different models that let you understand the interrelationships among the metrics.

Second, the value alignment system can utilize a variation of so-called trolley problems for supervised machine learning. A trolley problem is a thought experiment about a fictional situation in which you can save the lives of five people who'll otherwise be hit by a trolley by swerving and killing one person. Whether you choose to divert the trolley reveals your values. Variations of trolley problems change the number of people who die under each option and associate attributes with the people. They are pairwise comparisons. Trolley problems are useful for value elicitation because humans are more easily able to reason about small numbers than the long decimals that usually appear in trust metrics. Moreover, couching judgements in terms of an actual scenario helps people internalize the consequences of the decision and relate them to their use case.

As an example, consider the two scenarios shown in Figure 14. Which one do you prefer?
There is no right or wrong answer, but whatever you select indicates your values.

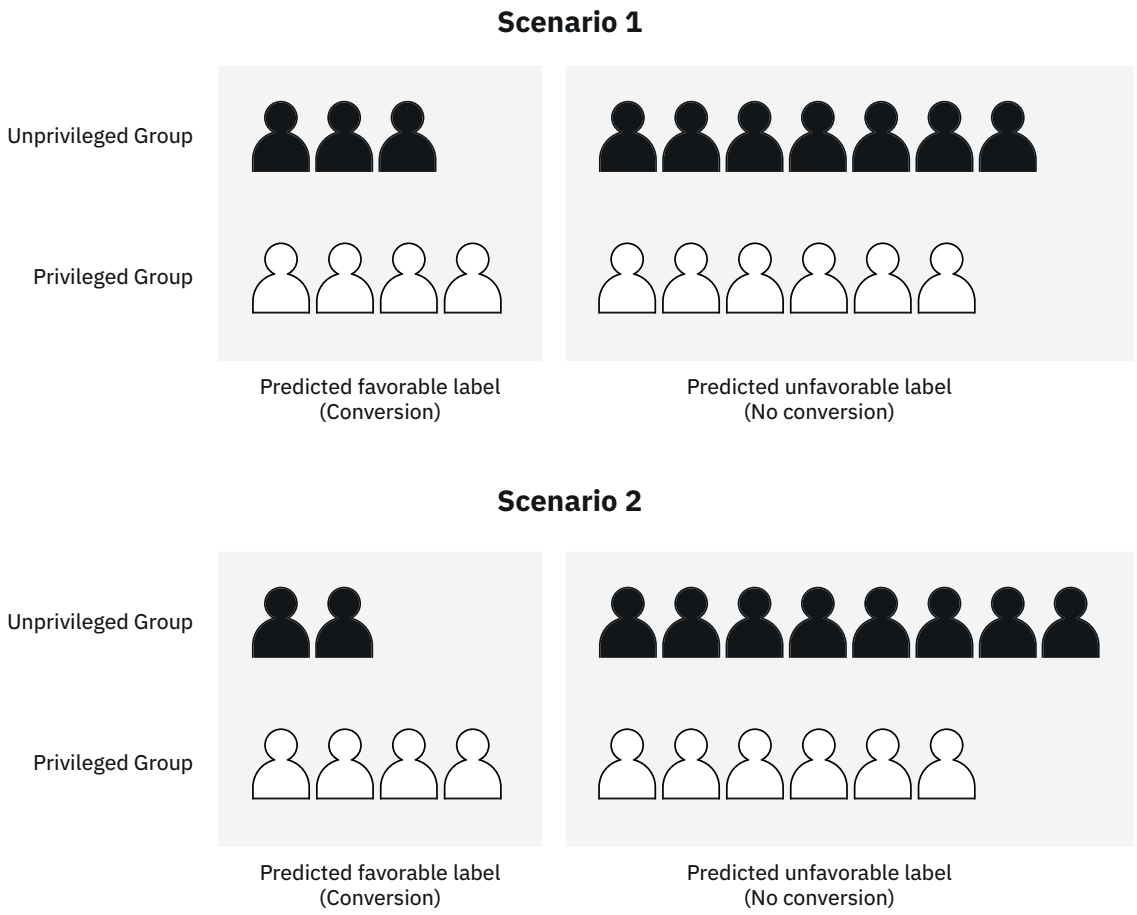


Figure 14: An example pairwise comparison between two scenarios that you can have a policymaker make a judgment upon.

Making these judgments is a business decision. The accuracy of an audience optimization model can be related to some business utility function related to the risk and benefit of serving ads to people who may or may not convert. Fairness is related to many considerations such as potential fines as part of regulations for high-risk AI, costs associated with a loss of brand reputation among customers, risk of legal claims, loss of morale among employees, etc. as well as the moral and ethical imperative to not cause undue harm to the most vulnerable people.

It is tempting to ask for universal best ranges of fairness metric values, but starting with such benchmarks is not fruitful. It really depends on the characteristics of the dataset and prediction problem, and the business costs and considerations. The only exception is if a law or regulation stipulates a fairness metric range, such as the four-fifths rule for disparate impact ratio. As an analogy, you do not usually ask for best universal ranges of accuracy either because that too depends on the problem at hand. For example, if the task is to predict heads or tails in a coin flip, any accuracy above 50% would be amazing. In bid optimization, an accuracy of 10-20% may meet industry standards. In the future, probabilistic id matching will have increasing minimum accuracy requirements. In the same way, you should consider the problem when determining what is an acceptable fairness metric value.

Fairness in Problems That Are Not Binary Classification

The vast vast majority of research and development in AI fairness has been on binary classification problems, with just a little work on ranking problems.

Multi-Category Classification

When dealing with classification problems that are not binary, but have many labels, the main concepts discussed in this document continue to apply. The confusion matrix gets expanded to the number of classes on each side, so that there are more individual errors than just false favorable rate and false unfavorable rate, but the concepts of independence, separation, and sufficiency remain the same along with their guidance having to do with worldviews and the presence or absence of social bias. You can take the ratios or differences of various individual errors as fairness metrics. For example, in a three category problem with labels 'conversion,' 'click,' and 'impression,' the rate of partial conversion difference across men and women could be a fairness metric.

Regression

Group fairness metrics for other settings, such as regression problems can be created by taking the difference or ratio of any existing predictive performance metric across groups delineated by protected attributes. The difference of mean squared error or mean absolute error across groups, would for example, be a reasonable fairness metric under the "what you see is what you get" worldview. Computing the maximum absolute difference (also known as total variation distance) or Kullback-Leibler divergence between the label distributions would be a reasonable fairness metric under the "we're all equal" worldview. These can be accomplished by binning the continuous response variables by quantiles. Otherwise, you can also convert regression problems into classification problems via thresholding the continuous response variable defined in a business-driven way. One bias mitigation algorithm in the open source AI Fairness 360 toolkit tackles regression problems in addition to classification problems. It is known as the grid search reduction bias mitigation algorithm. It is an in-processing algorithm.

Overall Flow

Given all the concepts discussed so far, what is then the overall recommended flow of the development lifecycle? The basic starting point is the life cycle shown in Figure 15 involving several stages from problem specification to deployment and monitoring along with the different roles involved. Let's go stage by stage and point to the recommended actions that go beyond what is typically done.

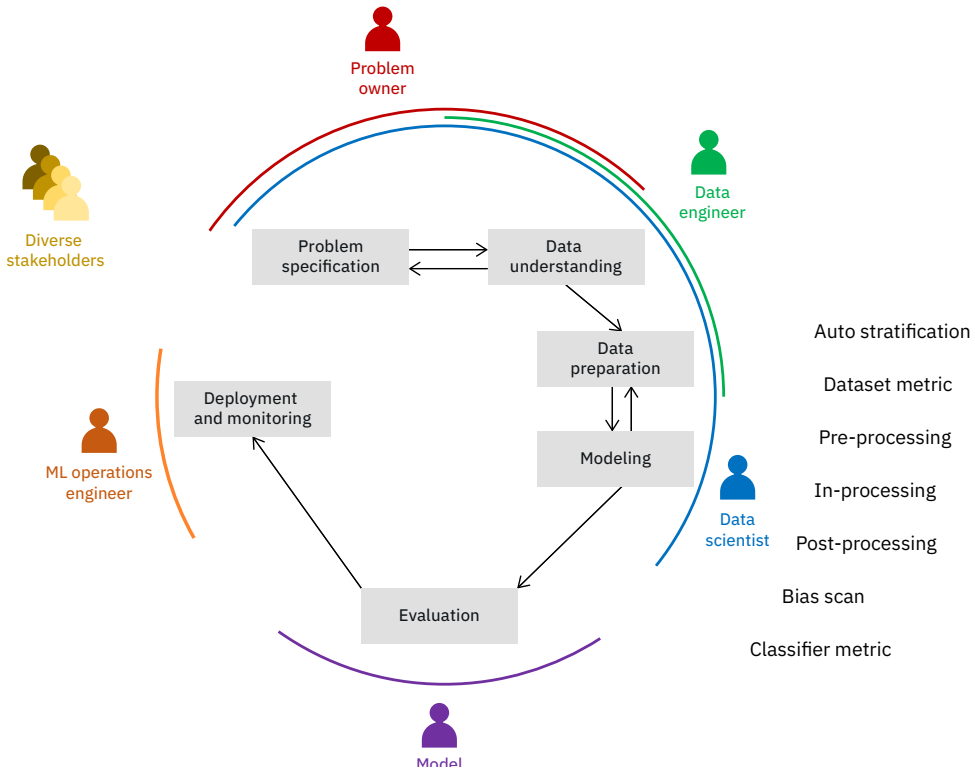


Figure 15: Typical machine learning model development lifecycle.

Problem Specification

In the problem specification phase, it is critical for the data scientists to collaborate with the problem owners (business and policy executives) to pin down whether a particular problem should even be worked on or not, and whether fairness is a top concern in that problem or not.

In this stage, the problem owners should also determine whether they are more interested in group or individual fairness as well as their worldview. Together these considerations will help determine the fairness metric of interest. Trolley problems to understand fairness metric ranges may be conducted at this point.

While teams that are set up for Fairness by Design practices will cover people from diverse backgrounds, it is encouraged to hear additional diverse voices, especially people from traditionally marginalized groups to inform the problem specification.

Data Understanding and Data Preparation

In the data understanding phase, the team should focus on understanding the potential sources of bias in measurement (social) and sampling (representation). This understanding will help inform the problem specification and worldview as well. In the data preparation phase,

care should be taken to not inadvertently introduce or exacerbate unwanted biases that may be present in the data. Exploratory data analysis on the prepared data should include multidimensional subset scanning to highlight anomalous subgroups and provide information to the problem owners. Once the protected attributes have been selected, you can start computing dataset fairness metrics.

Modeling

As part of the modeling pipeline, you should ensure that dataset and classifier fairness metrics are computed at the appropriate points and that bias mitigation algorithms are part of the pipeline as necessary. Creating several models and comparing them not only on accuracy, but on fairness and all the other relevant dimensions of trustworthiness, is recommended. A refinement of allowable ranges of metric values will become more apparent because it will be known at this point what is feasible and what is not.

Evaluation

Evaluating models by an independent model risk management team or model validator should be performed. Here too, it is important to study not only accuracy, but also fairness metrics and other metrics of trustworthy AI. Also in this stage, it is advantageous to bring in people with lived experience of marginalization.

Deployment and Monitoring

In the last stage of the life cycle, deployment and monitoring, you should set up monitors that observe the different trust metric values as they evolve over time and issue alerts if something starts to go awry. In addition, you should instrument your entire lifecycle to generate so-called facts that are captured and rendered as FactSheets as a means for transparency, AI governance and in support of compliance protocols.

For additional information on resources focused on the development of processes and practices to support the overall deployment lifecycle in marketing and advertising use-cases, explore "Understanding Bias in AI for Marketing, A Comprehensive Guide to Avoiding Negative Consequences with AI" published by the IAB AI Standards Working Group in 2021.

Metrics and Methods in the Advertising Toolkit for AIF 360

Throughout Part 2 of this document we have explored various bias identification metrics and mitigation algorithms adjacent to fairness concepts and their requirements. Below is a comprehensive list of all metrics and algorithms available within the Fairness 360 Toolkit. Both tables contain key information and application and further information including links to the academic papers, and python/scikit tooling, or original apis.

Table 3: Bias Mitigation Algorithms in AI Fairness 360.

Method	Description	Cat	Data	Fairness Concept	Protected Attributes Required?	Ref/ Papers
Reweighing	Weights the examples in each (group, label) combination differently to ensure fairness before classification.	pre	all	independence	no	link
Optimized preprocessing	Learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives.	pre	tabular	independence	no	link
Disparate Impact Remover	Edits feature values to increase group fairness while preserving rank-ordering within groups.	pre	tabular	independence	yes	link
Learning Fair Representation	Finds a latent representation which encodes the data well but obfuscates information about protected attributes.	pre	tabular	independence	yes	link
Adversarial Debiasing	Learns a classifier to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions.	in	tabular	independence, separation	yes	link
GerryFair Classifier	Algorithm for learning classifiers that are fair with respect to rich subgroups, where rich subgroups are defined by (linear) functions over the sensitive attributes.	in	tabular	subgroup independence, separation, sufficiency	no	link
Meta-Fair Classifier	Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized with respect to that fairness metric.	in	tabular	independence, separation, sufficiency	no	link
Prejudice Remover	Trains a classifier with a discrimination aware regularization term.	in	tabular	independence	no	link
Grid Search Reduction	Reduces fair classification/regression to a sequence of cost-sensitive classification/regression problems, returning the deterministic classifier/ regressor with the lowest empirical error subject to fair classification constraints among the candidates searched.	in	tabular	independence, separation	no	link

Exponentiated Gradient Reduction	Reduces fair classification to a sequence of cost-sensitive classification problems, returning a randomized classifier with the lowest empirical error subject to fairness.	in	tabular	independence, separation	no	link
Calibrated Equalized Odds Postprocessing	Optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective.	post	all	separation	yes	link
Equalized Odds Postprocessing	Solves a linear program to find probabilities with which to change output labels to optimize equalized odds.	post	all	separation	yes	link
Reject Option Classification	Gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a band around the decision boundary with highest uncertainty.	post	all	independence, separation	yes	link

Table 4: Fairness Metric Categorizations, Properties, and Synonyms.

Metric Names	Fairness	Bias	Assistive/ Non-Punitive	Best Value	Worst Value(s)	API Link
statistical parity difference, demographic parity difference,mean difference (as a binary label dataset metric)	independence	social bias, representation bias,data preparation bias	either	0	−1, 1	original API scikit-learn API
statistical parity difference, demographic parity difference, mean difference (as a classification metric)	independence	social bias, representation bias, data preparation bias	either	0	−1, 1	original API scikit-learn API
disparate impact, disparate impact ratio (as a binary label dataset metric)	independence	social bias, representation bias, data preparation bias	either	1	0, ∞	original API scikit-learn API
disparate impact, disparate impact ratio (as a classification metric)	independence	social bias, representation bias, data preparation bias	either	1	0, ∞	original API scikit-learn API
average odds difference, average difference in false positive rate and true positive rate, average difference in true negative rate and false negative rate	separation	~~social bias~~, representation bias, data preparation bias	assistive	0	−1, 1	original API scikit-learn API

average absolute odds difference, average odds error, average difference in absolute false positive rate and absolute true positive rate, average difference in absolute true negative rate and absolute false negative rate	separation	~~social bias~~, representation bias, data preparation bias	assistive	0	1	original API scikit-learn API
average predictive value difference, average difference in positive predictive value and false omission rate, average difference in precision and false omission rate, average difference in false discovery rate and negative predictive value	sufficiency	representation bias, data preparation bias	non-punitive	0	–1, 1	(need to construct) original API
average absolute predictive value difference, average difference in absolute positive predictive value and absolute false omission rate, average difference in absolute precision and absolute false omission rate, average difference in absolute false discovery rate and absolute negative predictive value	sufficiency	~~social bias~~, representation bias, data preparation bias	non-punitive	0	1	original API scikit-learn API
consistency, consistency score	individual	depends on choice of distance metric	depends on choice of distance metric	1	0	original API scikit-learn API
Theil index with $b_j = 1$ for true favorable labels, $b_j = 0$ for true unfavorable labels, $b_j = 1$ for false favorable labels, $b_j = 0$ for false unfavorable labels	individual and independence	social bias, representation bias, data preparation bias	either	0	1	scikit-learn API
Theil index with $b_j = 1$ for true favorable labels, $b_j = 1$ for true unfavorable labels, $b_j = 0$ for false favorable labels, $b_j = 0$ for false unfavorable labels	individual and separation	~~social bias~~, representation bias, data preparation bias	assistive	0	1	scikit-learn API

Part 3
Practical Applications and Examples:

As teams across the advertising ecosystem begin to experiment with the Fairness 360 toolkit and share their learnings, they can be featured below as illustrative applications and example notebooks that others can learn and experiment with. The intent is to provide practitioners with easy-to-follow scenarios to help explore adopting the metrics and mitigation strategies throughout their machine learning pipeline.

The underlying approaches used to design a system will greatly influence the appropriate measures in applying Fairness 360 tools. Not every example will be directly applicable to all parts of the ecosystem or even two similar offerings in the same space. For example, the “Fairness for Predictive DCO” below may not apply to all predictive models serving dynamic creative optimization.

Practitioners will be able to explore these notebooks and evaluate approaches and critical concepts to how the AI Fairness 360 metrics and mitigation tools are applied. Each overview below will outline the example, use case, and key considerations. Additional details are provided within the individual notebook examples and their related documentation.

Fairness for Predictive DCO

The Fairness for Predictive DCO (dynamic creative optimization) example here showcases how campaign data can be utilized to explore when model or data bias might impact the predictive DCO in a post-processing scenario.

Mitigate Bias by transforming the original dataset

The debiasing function we will use to mitigate bias is **Reject Option Classification (ROC)** -- the post-processing algorithm from AIF360. For this lets divide the dataset into training, validation and testing partitions.

In [16]:
```
# Split the standard dataset into train, test and validation
dataset_orig_train, dataset_orig_vt = advt_standard_dataset.split([0.7], shuffle=True)
dataset_orig_valid, dataset_orig_test = dataset_orig_vt.split([0.5], shuffle=True)
```

In [17]:
```
# print out some labels, names, etc.
display(Markdown("#### Training Dataset shape"))
print(dataset_orig_train.features.shape)
display(Markdown("#### Favorable and unfavorable labels"))
print(dataset_orig_train.favorable_label, dataset_orig_train.unfavorable_label)
display(Markdown("#### Protected attribute names"))
print(dataset_orig_train.protected_attribute_names)
display(Markdown("#### Privileged and unprivileged protected attribute values"))
print(dataset_orig_train.privileged_protected_attributes,
      dataset_orig_train.unprivileged_protected_attributes)
display(Markdown("#### Dataset feature names"))
print(dataset_orig_train.feature_names)
```

Training Dataset shape

```
(8309, 16)
```

Favorable and unfavorable labels

```
1.0 0.0
```

Protected attribute names

```
['age']
```

Privileged and unprivileged protected attribute values

```
[array([1.])] [array([0.])]
```

Dataset feature names

```
['age', 'gender=F', 'gender=M', 'gender=Unknown', 'income=<100K', 'income=>100K', 'income=Unknown', 'area=Rural', 'area=Unkn
own', 'area=Urban', 'college_educated=0', 'college_educated=1', 'homeowner=0', 'homeowner=1', 'parents=0', 'parents=1']
```

Metric for training data

In [18]:
```
metric_orig_train = BinaryLabelDatasetMetric(dataset_orig_train,
                                             unprivileged_groups=unprivileged_groups,
                                             privileged_groups=privileged_groups)
display(Markdown("#### Original training dataset"))
print("Disparate impact between unprivileged and privileged groups = %f" % metric_orig_train.disparate_impact())
```

Original training dataset

```
Disparate impact between unprivileged and privileged groups = 0.606221
```

Train classifier on original data

In [19]:
```
# Logistic regression classifier and predictions
scale_orig = StandardScaler()
X_train = scale_orig.fit_transform(dataset_orig_train.features)
y_train = dataset_orig_train.labels.ravel()

lmod = LogisticRegression()
lmod.fit(X_train, y_train)
y_train_pred = lmod.predict(X_train)

# positive class index
pos_ind = np.where(lmod.classes_ == dataset_orig_train.favorable_label)[0][0]
```

Objectives:
Study the application of fairness metrics to identify system biases in how predictive algorithms and segmentation affected the campaign's ability to engage consumers; example hypothesis:

Are there factors affecting the conversion rate, by conversion, we are referring to the click-through rate that is impeded by the model observing signals beyond general targetings like age ranges and DMA?

Datasets:
For this experiment the team focused on campaign data that utilized a combination of AI driven advertising tactics. The data-set was transformed from multiple log files derived from dynamic content optimization, which employed multiple logistic regression classifiers, to bid optimization and audience segmentation. In the case of this particular study, the model itself was not openly available, so the team utilized the highest winning predictive score available within the log files. This score was representative of the model's prediction of the likelihood a consumer would convert based on the creative served. This score was used to help determine the conversion rate predicted by the model.

Teams should allow for Extraction, Transformation and Load actions essential to remove noise and nulls within the data set while also translating key encoded information, for example gender, from code to "Male", "Female", or "Other" to assist in further analysis. Ambiguous data can be problematic, so it is important to evaluate and adapt data where possible.

KPIs & Feature Selection
It is important that before you proceed with a campaign data set, that you identify the campaign KPIs that you want to analyze for bias. Defining these features along with other features you might want to evaluate within the bias metric application are important and tied to how you transform the data.

Methodologies
The team utilized a novel new approach to detect attributes with bias called Automatic Stratification. It is a technique to find subgroups that are correlated to a particular output. This work efficiently scales stratification across multiple features simultaneously to identify the strata with the most unexpectedly high (or low) outcomes. This new approach enabled the team to explore different permutations of the privileged and unprivileged classes automatically rather than identifying each attribute using individual metrics, allowing for rapid exploration of the entire dataset.

In addition, Multi-Dimensional Subset Scan (MDSS) was employed to automate the identification of subgroups that have predictive bias.

Selection of privileged and unprivileged attributes for the bias study proved to be a bit of a challenge, predominantly because there is a higher propensity in advertising for many protected attributes to be present in the data. If we mitigate one bias it should not result in another attribute being affected by this mitigation. With the possibility of thousands of features across a dataset there was the potential for manual identification using metrics such as disparate impact and mean difference to become very time consuming. To solve for this and provide expedience the team created automation tools to detect these protected attributes.

Mitigation
Once the team successfully identified the presence of bias within the dataset they were able utilize the Reject Option Classification mitigation approach. ROC allows the application of favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups where the model's prediction is least certain (classification threshold). By exploiting this low confidence region to reduce discrimination, we can reduce the bias in the predictions.

Notebook
The Fairness for Predictive DCO notebook is provided as an example exploration for learning and exploration of the Advertising Toolkit for AI Fairness 360. You will find sample, synthetic data to utilize within the notebook, helping to deliver a simulated scenario with actual biased outputs and mitigation techniques.

Fairness for Audience Insights & Targeting in Media

Contributed by Mindshare

Objectives
Using individual-level customer data, we sought out to identify and evaluate bias and the impact it has on predicting sales for a CPG (Consumer Package Goods) brand. Bias can be identified in many places ranging from input data to algorithm selection to output data, as such, we structured our approach to evaluate each of these areas to understand how mitigation efforts across the entire modeling pipeline may result in different outcomes.

Datasets & Feature Selection
For our application we leveraged first party consumer CPG sales data, enriched with demographic and psychographic data to create a comprehensive view of those consumers. As part of our dataset, we had access to sales data from consumers who bought competitor products which allowed us to create the necessary flags to build a Classification predictive model.

With a dataset containing over 100 features, we evaluated data quality and inconsistency in a thoughtful way. Some of the imputation methods we applied included mean & mode application, others included variable transformation to reduce noise, and we also removed outlier observations. Also of note, our dataset contained a gender imbalance – indexing significantly higher for female consumers vs. male consumers. That being said, we decided to preserve this aspect (gender is a Protected Attribute – an attribute that partitions a population into groups whose outcomes have parity) of the dataset to better understand the impact of bias mitigation.

As part of our exploratory analysis, we employed feature reduction practices such as removing attributes that are highly concentrated in one value as these are unlikely to be key predictors. Additionally, we conducted a correlation analysis to understand attributes that may be redundant and were thoughtfully excluded where appropriate.

Methodologies
We evaluated a series of methods from the AI Fairness 360 toolkit to identify bias at each phase of the machine learning pipeline (e.g. pre-processing, in-processing, post-processing).

Pre-processing – generally, the most common source of bias can be found in the underlying data. Applying a pre-processing routine is a beneficial task to mitigating bias prior to training a model. At this layer in the pipeline, we chose to apply the Reweighing method which generates weights for the training examples in each (group, label) combination differently to ensure fairness before classification.

In-processing – learning algorithms are most impacted at this layer in the pipeline. We implemented an Exponentiated Gradient Reduction method that reduces misclassification and returns a randomized metric that helps quantify the lowest error within our fairness thresholds.

Post-processing – at this stage in the process we were unable to modify the input data or learning algorithm, as such, we implemented an Equalized Odds Postprocessing method that uses a probabilistic approach to balance how the output labels are categorized.

Mitigation & Results

Our approach consisted of building several bias-mitigated models and comparing the results vs. a biased model. We considered several evaluation and fairness metrics. In particular, we focused on the Disparate Impact fairness metric. Measuring Disparate Impact is a good way to prevent discriminatory practices that may inherently exist in model development.

In our study, Disparate Impact compares the proportion of individuals that are classified as male or female in our dataset that are assigned a specific outcome (buying a product). We found that Disparate Impact drastically improved by 6X when we mitigated bias and achieved optimal fairness, while also preserving similar prediction accuracy when compared to a biased model (<-1.8% variance). By implementing a bias-free model, we can identify male buyers that were previously unknown to us, improve our understanding of male buyers, and drive incremental marketing outcomes (e.g., sales) from this population while also preserving individual fairness across gender.

As mentioned above, our research shows the impact of mitigating bias does not materially degrade our model performance, and as a result, business outcomes can continue to be prioritized in a responsible way. In short, we can strike the right balance of fairness and performance, making advertising better for people and delivering 'Good Growth' for brands.

Part 4
Beyond Research

What is next?

Over the next several years, we see the advertising industry being reshaped into a fabric of human-aware platforms grounded in privacy, diversity, and transparency. Adjacent to the trust-oriented efforts in play right now, a proliferation of new channels for engagement may arise. Brands, agencies, and ad-tech platforms should seek to reflect on their current strategies and technologies and build a future that advances trustworthy AI with fairness, robustness, explainability, transparency, and privacy at the heart of their design.

As organizations explore the possibility of algorithmic bias within the advertising technology they employ they should also consider broadening the awareness beyond practitioners to the entirety of the team.
Revisiting a few steps from above:

- Organizations should provide all teams with Diversity, Equity and Inclusion training. The outcomes of these training sessions will prepare teams for the right mindset to question existing practices and platforms and prepare them for future developments.

- Explore cognitive biases and their cross-section with technological biases. Create a culture of questions around diverse points of view, helping to craft better strategies and develop better technology through an inclusive and understanding lens.

- Ask teams to invest time and energy in learning tools, like the AI Fairness 360 toolkit discussed above to help them assess existing platforms and augment the design, development, and maintenance of future efforts.

As the organization makes progress, key efforts should include developing an organizational approach to advocacy both internally, with partners, and across the industry. The only way for the evolution of our practices to truly take shape is through shared experiences, common dialogue and widely-accepted change. Brands, agencies and advertising technology organizations who have invested time and effort could consider sharing their findings and use-cases with the community of organizations taking part in this work.

If your organization has an interest in sharing findings or making use-cases or sample code available please contact: watsonadvertising@us.ibm.com

Additional Resources

While this document has focused predominately on the importance of and examples of how to explore biases in advertising technology and data, there is a plethora of available resources, tooling, guidance and education available to advance the industry's efforts towards DE&I and a fairer advertising industry. Below is a collection of resource for consideration:

IBM
IBM Research is working on a range of approaches to ensure that AI systems built in the future are fair, robust, explainable, account, and align with the values of the society they're designed for. IBM is working to ensure that in the future, AI applications are as fair as they are efficient across their entire lifecycle. IBM has also developed widely accepted practices for ethics by design, fairness by design, and security and privacy by design — to ensure that systems are built from the ground up with humans at the center of their strategy.

- AI Ethics:
  https://www.ibm.com/artificial-intelligence/ethics

- Fairness, Accountability & Transparency:
  https://research.ibm.com/teams/fairness-accountability-transparency

- Adversarial Robustness & Privacy:
  https://research.ibm.com/teams/adversarial-robustness-and-privacy

- Ethics by Design for AI:
  https://www.ibm.com/design/ai/ethics/

- Advancing AI Ethics Beyond Compliance:
  https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics

4As

The 4As serves as a continuous voice in helping agencies transform toward human-centered approaches. Their Campaign Enlightenment program helps agencies and brands explore how to eliminate bias and create authentic content, examine existing agency processes and apply practical tactics, and make changes that enhance the work the agency puts into the world. In addition, the 4As have led extensive research and thought leadership on the topics of diversity and inclusion.

- Explore the Campaign Enlightenment Program:
  https://learninginstitute.aaaa.org/library/info-course-campaign-enlightenment-program-178262/398268/about/

- All DE&I Programming:
  https://learninginstitute.aaaa.org/library/?category=Diversity%25%3B+Equity%2C+and+Inclusion

- Upcoming DE&I programming, including live workshops:
  https://www.aaaa.org/learning-institute/

IAB

The Inclusion Institute is an IAB initiative committed to realizing a truly inclusive, equitable and diverse industry that is sustainable. IAB is seeking to accomplish this through outreach, education, leadership development and accountability done with our partnerships in the digital media ecosystem. The IAB AI Working Group examined the potential for bias in digital marketing and advertising producing a thorough and in-depth primer to avoiding negative consequences with AI.

- Explore the Inclusion Institute:
  https://www.iab.com/organizations/inclusion-institute/

- Read IAB's AI Standards Working Group's
  "Understanding Bias in AI for Marketing, A Comprehensive Guide to Avoiding Negative Consequences with AI".

ANA's SeeHer
To help benchmark success, in 2016 SeeHer spearheaded the development of the **Gender Equality Measure (GEM®)**, the first research methodology that quantifies gender bias in ads and programming. GEM® is an award winning methodology and has quickly become the global measurement standard, measuring 200,000+ ads, representing 87 percent of worldwide ad spend. The SeeHer Marketing Essentials Toolkit, is a comprehensive turn-key set of data, insights and tools to help create a more gender equal society and more effective
business ecosystem.

- SeeHer Marketing Essentials Toolkit:
  https://www.seeher.com/insights-and-tools/seeher-marketing-essentials-toolkit/

- SeeHer GEM
  https://www.seeher.com/what-is-gem/

Female Quotient
The Female Quotient is an equality services company that provides thought leadership platforms to women, and develops solutions for organizations committed to closing the gender gaps at work. It is fueled by the ideas, ambitions, innovations, and empathy of working women around the world. It is a combination of live events, online forums, custom research, media, and advisory services. It is a collective enterprise led by women, and welcomes the insights of everyone engaged in the issue. It identifies challenges, surfaces effective strategies, forges powerful networks, and ultimately sparks real progress.

- Equality Lounge® at Industry Events

- The Business of Equality
  There are many agencies also engaging in dialogue and innovation to craft a future where our practices, processes and outcomes are empathetic, and intention driven.

Mindshare
Mindshare has driven client education and thought leadership with their Good Growth positioning to create a more enduring, diversified and sustainable world through more empathetic, inclusive marketing solutions. This includes their "On the Basis of Code" work, an inclusive innovation practice driving intentional marketing practices, including the industry-first inclusion PMP series in direct response to pervasive biases in ad-tech and the GroupM-partnered Data Ethics solutions.

Acknowledgments

This document was developed by IBM Watson Advertising with contributions from IBM Research, advertising industry organizations and agencies. Mindshare Media is an early adopter of the Advertising Toolkit for AI Fairness 360 and their work in exploring bias in their tools and approaches is available in section 3. IBM Watson Advertising would like to thank the following organizations for their support, advocacy and continued focus on critical issues helping to transform our industry:

Mindshare Media
WPP
GroupM
The American Association of Advertising Agencies
Internet Advertising Bureau
The Association of National Advertisers
Female Quotient

Produced in the United States of America
June 2022