Deloitte.

Striving for fairness in AI models

Executive Summary
In this paper, we address algorithmic biases, especially those that may result in unfair and discriminatory practices. Unfair bias prevents us from becoming our better selves, it bears economic productivity costs, and holds back the advancement of science and society in general. Much has been achieved over the past decades to bring the many pernicious forms of bias to light, to combat it. Yet much more remains to be done. The more obvious cases are being tackled by lawmakers and society. The more subtle – yet no less harmful – forms continue to exist, hidden from view. Our role, as responsible data scientists, is to root it out from the less obvious places, to ensure our analyses are objective and fair, that we make decisions based on the right data for the right reasons.

Fortunately, with the right combination of awareness, governance, and analytical tools, bias can indeed be effectively managed. First and foremost is to sensitize staff for the red flags and the dangers of bias. Second is to elevate fairness to become an optimization criterion, next to model performance. Multi-dimensional optimization is less straightforward, but nothing that seasoned data scientists can't handle. Third, staff must be given the proper incentives and tools to analyze bias, just as they are given to optimize performance. Lastly, organizations must ensure bias – and ethics in general – are a boardroom topic operationalized by appropriate governance routines.

MAKING AN IMPACT THAT MATTERS since 1845

Confronting bias is now more important than ever

Bias is everywhere. It can be found in society, such as the skewed distribution of life-expectancies, incomes or population densities. Or in commerce, such as the distribution of credit defaults or real estate prices. Bias does harm, and everyone knows it. In the form of racial prejudice, it has subjected millions to lack of social and economic opportunity for generations, leaving them vulnerable to exploitation. It erects often insurmountable barriers to the underprivileged, some of whom, given the chance could have contributed substantially to the advancement of our civilization. Bias in the form of sexism has relegated women to traditional roles, even if they may have been far more capable than their male counterparts. Throughout history, bias in the form of religious intolerance has fomented violence against those of other beliefs.

Ironically, process digitization does not inherently render decisions more objective. To the contrary, biases and inequalities are exacerbated when cemented into easily scalable decision engines. Algorithms can predict who will default on a loan, who should be hired, and what price each customer is willing to pay for a product or service. They can learn much more from data than humans can digest, identifying patterns in the predictions in unexpected ways. Those patterns are sometimes associated with who we are, including our race and gender. More concerningly, they can reflect past discriminatory and/or exclusionary practices.

This realization weighs heavily on business leaders who want neither the company nor personal brands to be associated with systematic unfair discrimination. Machine learning algorithms are increasingly driving high-impact decisions across multiple industries. Concerns about perpetuation of bias have led scholars worldwide to introduced numerous definitions of fairness and their corresponding mathematical formalizations. That may make fairness seem more objective, but that would be too hasty a conclusion.

A mathematical perspective on bias

No single algorithm can satisfy all the various definitions of fairness, many of which are mathematically incompatible with each other. Choosing one means foregoing another. Selecting a fairness definition is in itself problematic because fairness is not binary nor absolute. What fits to one situation will not in another. (While quite a current topic for AI, this debate about fairness is nothing new: It has been debated among philosophers for millennia, from aristotle to rawls.) There is no more consensus beyond ivory towers and philosopher's caves. Consumers disagree vehemently about what it means to be fair, further complicating matters for practitioners. Many popular notions of fairness assume a clear distinction between "legitimate" features (e.g. income) and "irrelevant" features (e.g. race). Yet these overlook proxies of the prediction, which can be closely correlated with proxies of personal characteristics. Existing fairness definitions fail to address discrimination already embedded in the data.

To address this, we will seek not only to identify model design decisions impacting fairness, but also investigate why those biases may exist. In our mission to stamp out bias, we must first thoroughly understand it, which requires analysis from multiple perspectives.

01. Protected feature impact analysis: Quantify the impact to fairness associated with each of the protected features. Determine whether the model is generally at risk of unfairly discriminating.

02. Protected group risk assessment: Dive deeper to explore which specific protected groups are at high risk of unfair discrimination.

03. Non-protected feature risk assessment: Investigate the non-protected features in the system as additional potential sources of bias. Explore which data points might drive the model to discriminate unfairly against protected groups.

04. Model assessment and tracking: Examine the trade-off between performance vs fairness for multiple systems (or iterations of the same system) to make a balanced decision on which to deploy.

> Selecting a fairness definition is in itself problematic because fairness is not binary nor absolute. What fits to one situation will not in another.

The first step in protected feature impact analysis is the understanding of the data set in the context of fairness. Here, we answer key questions surrounding bias: How are different groups represented in the data? Are there systematically different outcomes for different groups?

The central risk of bias pertaining to machine learning (M) is the bias inherent within the training data. What sets ML models apart from other types of models is that they derive their prediction or decision rules from training data, as opposed to being given explicit rule-based instructions. The orientation around data is what makes machine learning so powerful. It is also a vulnerability. The general rule "garbage in, garbage out" is particularly relevant for machine learning. Specific to our case, the risk is "bias in, bias out." To begin any analysis, we must collect the appropriate data.

This is best illustrated by example. We make use of US mortgage application data, as it is particularly rich in demographic data. We will continue to use this example dataset for illustration purposes throughout the rest of the analysis. The data we need to collect:

Tab. 1 – Data required for a quantitative analysis of bias

Data type	Example "US mortgage applications"
1. Protected features	... Age, race, sex, ethnicity
2. Non-protected features	... Income, credit score, requested loan amount
3. Ground truth and/or model predictions	... Applicant approved/denied the loan

Alongside data, we must define a specific business goal. In the case of bias detection, we must first define precisely what we mean by bias, a paradigm against which to measure the models (and their iterations). We explain different methods to measure bias using the example of a bank using AI to decide whether or not an applicant shall be granted a loan. In this example, the terminology "favorable outcome" indicates the applicant is granted a loan, a "correct favorable outcome" indicates that granting the loan was a good decision – namely, the loan is eventually repaid to the bank (the business goal).

Tab. 2 – Various measures of bias, explained

Outcome disparity	The proportion of favorable outcomes for a group with a particular protected characteristic value (e.g., Asian) is the same as the proportion of the entire population (i.e., everyone in the dataset) having a favorable outcome. E.g., the proportion of Asians granted a loan is the same as that for all loan applicants on average.
Statistical parity difference	The proportion of favorable outcomes for a group with a particular protected characteristic value (e.g., Asian) is the same as the proportion of other groups (i.e., everyone other than Asians) having a favorable outcome. E.g., the proportion of Asians granted a loan is the same as for other ethnicities.
Equal opportunity difference	The proportion of correct favorable outcomes for a group with a particular protected characteristic is the same as the proportion of correct favorable outcomes for the overall population. E.g., the proportion of Asian borrowers with a particular protected feature repaying their loan is the same as the proportion of the whole borrower population repaying their loans.
Average odds difference	The proportion of incorrect favorable outcomes and the proportion of correct favorable outcomes is the same between a group with a protected characteristic value and the entire population, e.g., the proportion of Asian borrowers who don't repay their loans is the same as the whole population and the proportion of Asian borrowers that do repay their loan is also the same as the whole population.
Disparate impact	The proportion of the group with a protected feature getting a favorable outcome is the same as the group having a particular protected characteristic value with the highest proportion of favorable outcomes, e.g., African Americans get offered a loan as frequently as White Angle Saxon applicants (assuming White Anglo Saxon is the most favored race characteristic for getting loans).

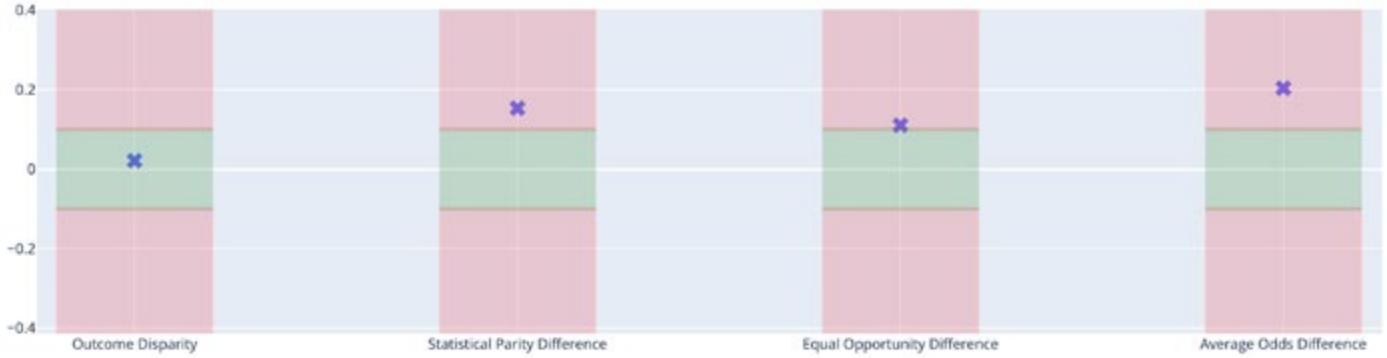
Each of these metrics represents a mathematical process for calculating how a certain protected group is being treated. Outcome disparity, for example, measures the difference in likelihood of a favorable result for a given protected group vs likelihood of a favorable result for the average person. Following the US mortgage applicant dataset, if an average applicant has a 50 percent of approval, but women on average only 40 percent, then the protected group women would have a negative outcome disparity value, 40 percent - 50 percent = -10 percent.

We methodically calculate the respective fairness metric for each group and then aggregate to gain insight into the overall fairness of the system. It is important to note at this stage that the choice of fairness metric will be driven by the context of the problem we are trying to solve. The goal of achieving no statistical parity difference between groups is not the same as achieving no equal opportunity difference. The former may require men and women to have similar loan approval rates, while the latter indicates that among creditworthy individuals, men and women have an equal chance of approval.

A side-by-side comparison provides an effective illustration of the difference in fairness metrics. The graph shows four metrics for the protected group of race. In there is a green band in which the group would be being treated fairly. (In this example the "fair" zone has been set to +/-10%.) Outside of this fair zone, the group is treated unfairly, whether favored (above the green band) or disadvantaged (below). The ethnicity "White", represented by a blue cross, which – in the case of US mortgage applications at least – is treated fairly for outcome disparity, but unfairly favored against other groups for the remaining three metrics.

The choice of metric is delicate and case-specific: It is entirely possible that these fairness metrics may contradict each other, some indicating fairness where others do not. Careful consideration must be given to this choice of which metric is the most appropriate.

Fig. 1 – Four different metrics for the race protected group



Source: Model Guardian

Protected group risk assessment

To begin with the fairness analysis, we first check the high-level impact on fairness experienced by each protected feature in the system.

The x axis measures the mean outcome disparity (or whichever fairness metric was being used) for each protected feature. This value is calculated as the mean of the calculated fairness metric for each protected group of a certain feature, weighted by that group's population size. In this way we can see to what extent the average applicant's outcome in the system is impacted by each of the other protected characteristics.

The y axis measures the maximum outcome disparity for each protected feature. This value is calculated as the maximum outcome disparity for all the subgroups of a certain feature. This value will therefore capture protected groups in the system that might be poorly represented, a minority group, and face significant disadvantage, e.g., low loan approval rates compared to other groups.

Based on these two metrics, we define four categories within which protected features can fall.

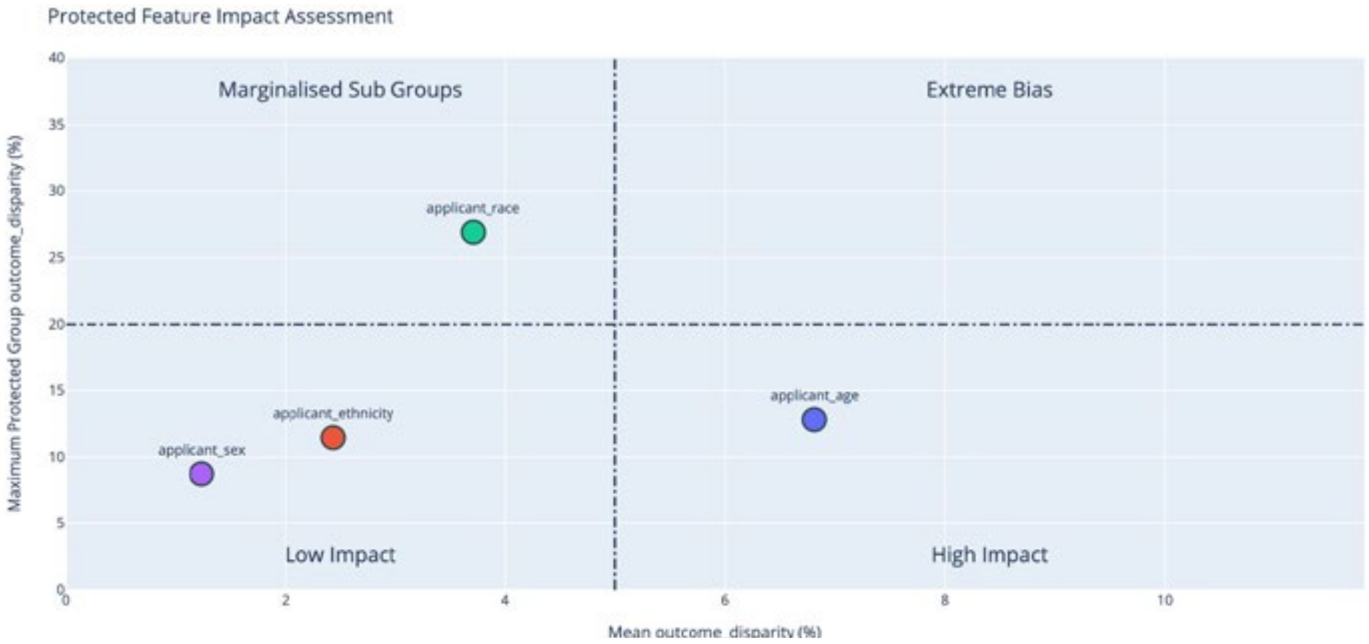
01. Low impact: In this example we see applicant_ethnicity and applicant_sex fall within this category. These features neither have a significant impact on the average applicant nor are there any protected subgroups within these features that see high levels of discrimination.

02. High impact: In this example we see applicant_age falls within this category. High impact features are protected features which are having a significant impact on the fairness in the system for the average applicant.

03. Marginalized subgroups: In this example applicant_race falls within this category. These features are not impacting fairness for the average applicant but there are particular protected subgroups which are at high risk of being treated unfairly.

04. Extreme bias: Thankfully in this example there are no features which fall within this category. Features found here would be driving high levels of bias not only for the average applicant, but also for underrepresented minority groups, and so would be of the highest concern.

Fig. 2 – Model features assessed by outcome disparity, categorized into bias risk groups (quadrants)



Source: Model Guardian

From this graph we can quickly gain two key insights.
First, there is bias in our system. If it was the case that all of the protected features in the system fell within our low impact category, we could conclude that the level of bias was minimal and acceptable. However it is the case that age and race fall outside of these thresholds and so we can conclude that our system is treating people unfairly.

Second, we know how to best further explore these features. The applicant's age has an impact on the average applicant, so the total population should be analyzed in order to figure out how this bias might possibly be remedied. The applicant's race on the other hand is problematic for specific minority groups, and so further investigation should be conducted to ascertain who these minority groups are and what specific remedies are appropriate, whether technical or non-technical. For example, if the bias is due to under-representation, collecting representative data or over-sampling and data balancing techniques would be appropriate. If the bias is due to subconscious biases of data labellers, employee training would better address the issue at its source.

De-biasing (pre-/in-/post-processing) methods are a subject of some debate, deserving of mention here. De-biasing rests on the premise that the undesirable bias or inequality can be measured and easily separated from legitimate or acceptable biases or inequalities. It proposes to surgically remove that undesirable bias from the data and from the model. Where this is theoretically possible, in practice it is seldom feasible because of the complexity in distinguishing legitimate drivers of the outcome (e.g. credit risk) from those affected by demographic identities (e.g. gender/race). Performing the surgery incorrectly would compromise a model's accuracy. Recent studies (e.g. "Delayed impact of fair ML") have even found de-biasing algorithms may harm the very people they seek to protect, by skewing their risk profile over the longer term.[1]

Ultimately, a thorough mitigation strategy should tackle bias at its source, embedded into the process that originally caused the bias, rather than mathematically neutralized. Bandage solutions will only reduce incentives to properly mend the underlying broken process. The exact approach will depend on the source of bias. For example, if staff assigned to identify social media posts as harmful are biased against certain dialects, their labeling may suffer a bias. Technical mitigation would be less effective than training – or a more diverse staff.

That is not to say that analytical methods do not have their place in de-biasing. Rather than using them to isolate bias for surgical removal, we argue they are better put to use in identifying the potential sources for subsequent process improvement.

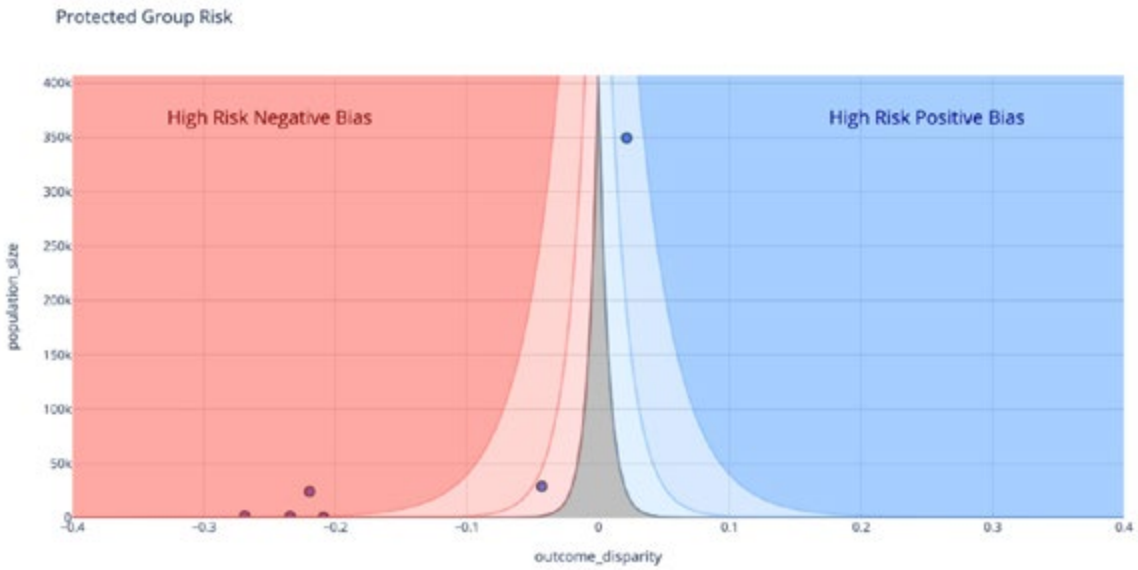
De-biasing methods are better applied toward identifying root causes than attempting to surgically remove bias, which can introduce unintended effects.

[1] i. Liu Lydia T, et al "Delayed impact of fair machine learning." International Conference on Machine Learning. PMLR, 2018.

Protected feature impact analysis

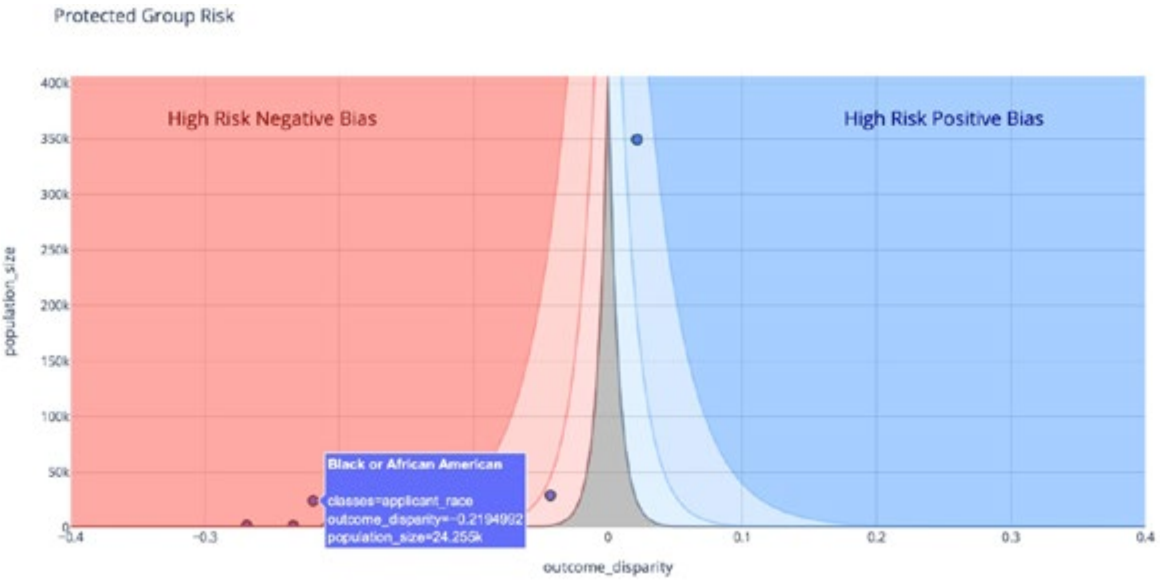
Following the line of investigation from the previous section, we dive deeper in order to pinpoint which specific protected subgroups of race were the most at risk in our system.

Fig. 3 – Representation of the risk level for different protected groups of being treated unfairly



Source: Model Guardian

Fig. 4 – Representation of the risk level for different protected groups of being treated unfairly with a more granular view on a protected group



Black or African American

classes=applicant_race
outcome_disparity=-0.2194902
population_size=24.255k

Source: Model Guardian

Each point on the chart is a protected group. In this case we have isolated all subgroups of race: White, Asian, Black or African American, American Indian, Native Hawaiian and other.

The x axis of this graph is the value of outcome disparity within the system. As mentioned previously, a perfectly fair value of outcome disparity corresponds to 0 percent. This would indicate the group was being treated consistently with the general population. As outcome disparity values increase (blue zones on the chart) the group tends to unfairly benefit from favoritism in the system. Groups in these zones are likely to more frequently experience positive outcomes.

Increasingly negative values of outcome disparity (red zones) represent an increasing risk of unfair and unfavorable discrimination of the groups by the system. Groups in these zones will likely experience a positive outcome less frequently than other groups.

The y axis of this graph depicts the size of the protected groups population, meaning how well they are represented in the system. In this example, the majority of groups have fewer than 50,000 members in the system, while the majority group (race = White) is represented with over 350,000 members.

We can calculate a risk score for each group based on the two parameters Outcome disparity and Population Size. As a general rule, we can consider that risk will correlate with size and level of unfairness found in the protected group: the larger the group and the higher the level of unfairness, the greater the risk. In this example, we deduce the following groups are high risk:

01. Black or African American. With an outcome disparity of -22 percent

02. Native Hawaiian. With an outcome disparity of -23 percent

03. American Indian. With an Outcome disparity of -27 percent

The conclusion of the outcome disparity analysis is that American Indians are the worst affected with a 27 percent lower chance of experiencing a positive outcome compared to the average applicant.

Non-protected feature risk assessment

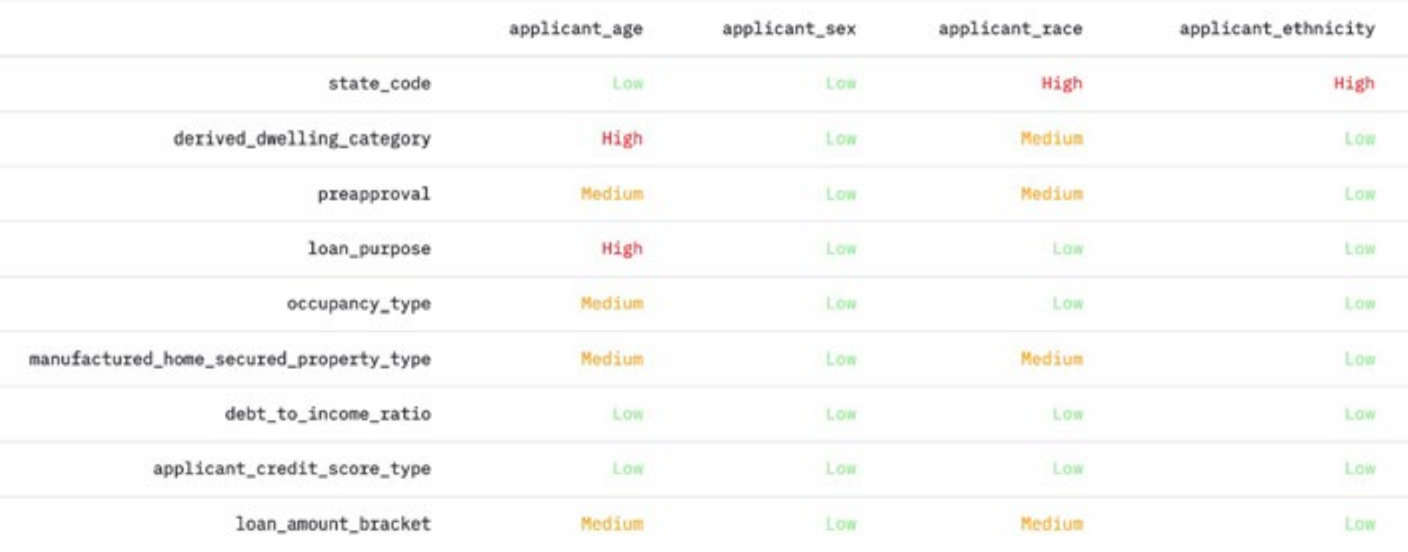
We have identified whether bias is present within our system, the extent to which that bias is an issue, and who is impacted. The next step in our assessment process is to examine why the system is biased, what potential sources attribute to the discrimination we have observed.

Most regulations block the use of protected features within an application or other selection process. Unfortunately, that has only limited effect in removing bias. Instead, bias finds its way through proxy variables into a machine learning system. Proxy variables are model features which do not directly represent a protected characteristic, but do correlate highly with a certain protected characteristic – so much so that a complex model can learn to infer protected information from this feature.

The US mortgage application example provides a useful illustration of this effect. The feature profession identifies some applicants as construction workers, others as nurses. Statistically, construction workers are overwhelmingly mail and nurses more often female. By including profession in the dataset, we inadvertently introduce a proxy variable which allows the model to infer the gender of applicants and thereby revert to historical gender discrimination within the training set. Neither model designers nor the AI technology itself are intentionally introducing bias, and yet it can enter into the model and effectively perpetuate historical prejudice – at scale – in future, AI-enabled and automated decision processes, such mortgage loan acceptance.

The most effective means to address this risk is to analyze each relationship between the protected features and the non-protected features, measure the strength of correlation for each one of these relationships and based on this strength of correlation assign a risk level.

Fig. 5 – Proxy risk matrix for a subset of non-protected features in our demo dataset



	applicant_age	applicant_sex	applicant_race	applicant_ethnicity
state_code	Low	Low	High	High
derived_dwelling_category	High	Low	Medium	Low
preapproval	Medium	Low	Medium	Low
loan_purpose	High	Low	Low	Low
occupancy_type	Medium	Low	Low	Low
manufactured_home_secured_property_type	Medium	Low	Medium	Low
debt_to_income_ratio	Low	Low	Low	Low
applicant_credit_score_type	Low	Low	Low	Low
loan_amount_bracket	Medium	Low	Medium	Low

Source: Model Guardian

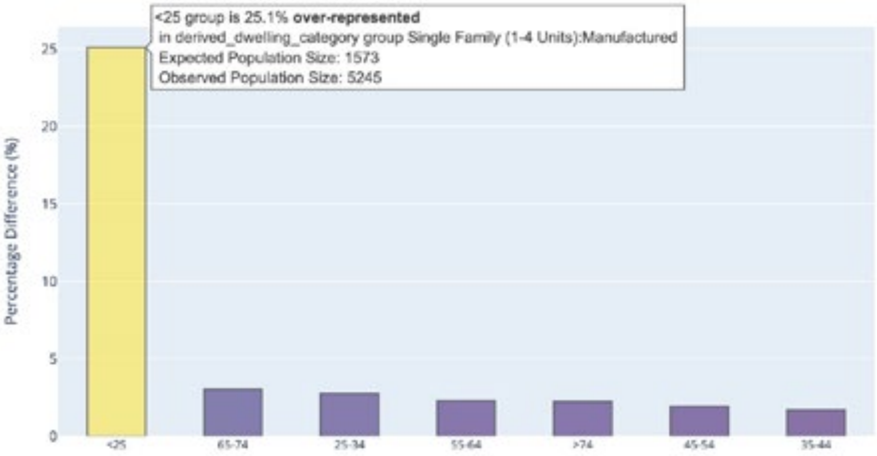
Protected features (age, sex, race and ethnicity) form the column headers and the non-protected features (loan purpose, debt to income ratio, applicant credit score type) the rows. A simple correlation (low, medium, high) will suffice to characterize the relationships between protected and non-protected features, giving a quick indication which of these may be proxy variables.

In the mortgage example, we see the relationship between an applicant's age and the derived dwelling category (summary type of property sought) show a high correlation, hence is a likely proxy variable. By including derived dwelling category as a feature into model, we may inadvertently introduce unintended bias against certain age groups.

The proxy relationship between age and dwelling category can be analyzed in more detail to determine which groups are at most risk of being affected. The proxy analysis shows as a histogram the extent to which each subgroup of age is correlated with dwelling category. By far the most affected group are people below 25 years old. This group is also 25.1 percent over-represented in the dwelling category: single family manufactured homes. A model using dwelling category to predict loan outcome may inadvertently introduce biases based on applicant's age, even if age is not explicitly included in the input data. Algorithms can predict who will default on a loan, who should be hired, and what price each customer is willing to pay for a product or service. They can learn much more from data than humans can digest, identifying patterns in the predictions in unexpected ways. Those patterns are sometimes associated with who we are, including our race and gender. More concerningly, they can reflect past discriminatory and/or exclusionary practices.

Academics have found that "fairness through unawareness" by excluding protected characteristics is largely ineffective, especially where other "legitimate" features encode this information or act as a proxy. See: "Fairness through awareness."

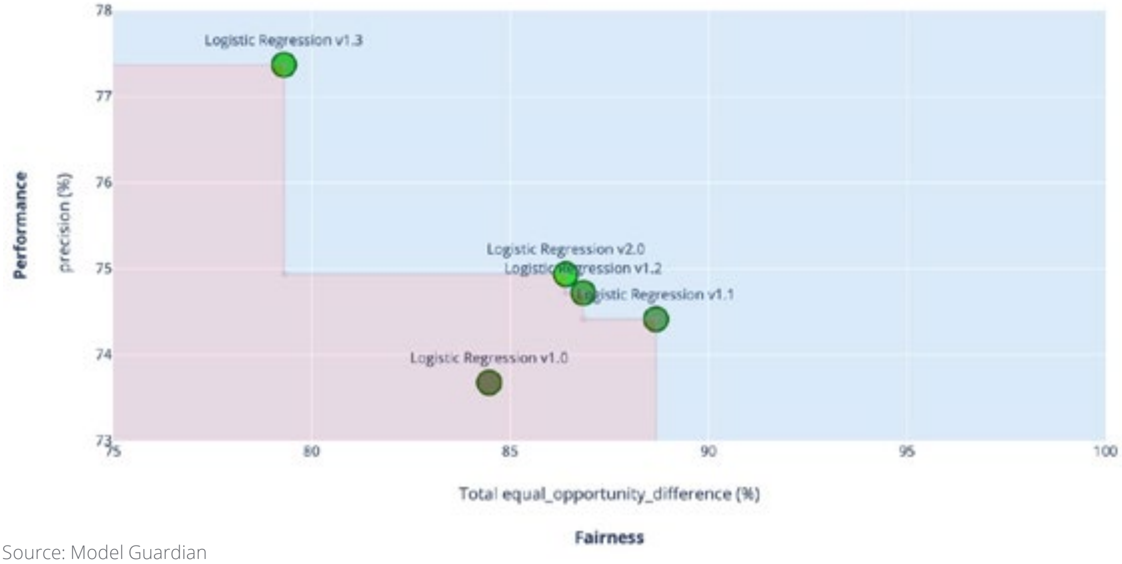
Fig. 6 – Proxy analysis – detailed view



Source: Model Guardian

Fairness by excluding protected characteristics is largely ineffective especially where other "legitimate" features encode this information or act as a proxy.

Fig. 7 – Model fairness vs performance trade-off comparison



Source: Model Guardian

Model assessment and tracking

We have so far analyzed features of a particular model or dataset. Our objective goes beyond identifying bias, however. We aim to improve models by designing them to be less susceptible to bias. The process of a refining a machine learning model is generally an iterative one – whether optimizing to performance metrics (accuracy, precision, recall…) alone, to maximize fairness objectives, or to balance between both. Model developers need to understand how changes to their models manifest themselves in outcomes – not along one dimension of performance nor of fairness alone, but in evaluating both simultaneously. This provides us a more sophisticated and holistic means to measure the relative efficacy of models – no longer "as accurate as possible", but rather "as accurate as possible whilst maximizing fairness."

The two dimensional evaluation shows the trade-off in performance (y axis, in this case precision) vs fairness (x axis) for five versions of a simple logistic regression model trained on the example US mortgage applicant dataset. The training data for each model has been curated slightly to produce a different outcome. In one case, certain groups have been given greater representation in the dataset to provide a more balanced training dataset. In another, certain high risk proxy variables were transformed or removed entirely to reduce the risk of bias in the model.

The results are eye-opening. We clearly detect the performance-fairness trade-off for each model. Performing this analysis for many iterations, we observe an "efficient frontier" where the model achieves a maximum fairness for that degree of precision (discriminatory power). Any model that lies along the efficient frontier is an acceptable iteration – from a technical standpoint. The choice of model will then come down to the priorities of the designers – whether greater precision or greater fairness is desirable. A model selection trade-off decision could be driven by internal policies, for example a team is trying to maximize their systems fairness but doesn't want to compromise on performance below a certain threshold. Or the decision could be externally driven, for example a regulation could mandate that a production model was trained to have no more than a certain amount of outcome disparity between protected groups.

In contrast, the logistic regression v1.0 model is sub-optimal, below the efficient frontier in the red-highlighted zone. (The frontier represents the maximum performance, so no models can lie in above the frontier in the blue area of the graph.) The mapping of the models along the performance-fairness plane allows developers to quickly prioritize their efforts, abandoning any models, such as logistic regression v1.0 and to concentrate on tuning one of the others.

Guarding against bias

Data scientists have the fortune of choosing between many fairness methodologies and toolsets, many of which are available as open-source. The tools differ widely in approach and, as research has found, in depth or suitability for different data and algorithmic settings.

Steep learning curve required to use the toolkits and limited guidance on metric selection

01. Imbalance between information overload vs over-simplification of complex results

02. Need for "translation" for a non-technical audience

03. Limited accessibility of toolkit search process

04. Limited coverage of the model pipeline

05. Limited information on possible mitigation strategies

06. Limited adaptability of existing toolkits to a customized use case

07. Challenges in integrating the toolkit into an existing model pipeline

This leaves significant gaps in the existing open source fairness toolkit landscape, which led us to develop our own tool, taking the principles and approaches discussed – as well as interactions with our many clients, and investing them into our own tool in order to improve the quality of bias analysis performed by Deloitte practitioners around the world.

A rigorous analysis of bias in machine learning models is a demanding exercise and fraught with inconsistencies in approach between practitioners. For these reasons, the Deloitte aiStudio invested the learnings and expertise amassed over years of research into a guided analytical tool, which we aptly named Model guardian.

The tool provides model builders and business owners alike the ability to detect potential unfair bias, investigate its source, and then to monitor progressive iterations of AI systems for bias as well as for effective predictive power. It examines how and to what extent the system is biased, who is at risk of being discriminated, and why the bias may exist.

Fig. 8 – A methodical approach to treating bias

01 **Identify bias in your dataset on a variety of fairness metrics**
Mathematical definitions of fairness cannot all be simultaneously met. Discover in which fashion your model is biased along any dimension.

02 **Investigation why these biases exist**
Quantitative analyses of potential proxies to protected features along with an assessment questionnaire to identify biases along the lifecycle.

03 **Optimize models for both performance and fairness**
Explore the trade-offs between the key fairness indicators (KKFI) and key performance indicators (KPI).

04 **Compare KPI/KFI across models**
Compare iterations of the same model to track performance (KPI) vs fairness (KFI).

05 **Communicate findings**
Automated reporting with key bias risks flagged from the quantitative and qualitative assessments.

Model guardian can be accessed either through the graphical user interface, which provides a guided workflow, or directly through invoking the python library, through which its functionality can be embedded directly into the models it is investigating. It connects seamlessly into the Deloitte AI ethics scorecard forming a comprehensive AI ethics assessment of the model. For each application, subject matter experts across ethics, AI, regulation, and digital risk management work to customize the assessment and provide a gap analysis of existing processes and controls.

Perhaps the greatest strengths of model guardian are the combination of features on offer: Its breadth throughout the model build pipeline, its highly customizable architecture, its logical workflow and intuitive visualizations – opening up the analysis of bias to a wider set of business stakeholders. Developers particularly appreciate its functionality to continuously monitor key fairness and performance metrics with progressive iterations of their models and data sets.

Model guardian can ingest both training datasets and model predictions. In this way the tool can be used during the model building process to analyze the potential ethical quality of training data, or it can be used once a model has been trained to analyze that particular model's performance and fairness. The tool analyzes relationships between the protected features and the non-protected features and assigns a risk-grading depending on the degree of correlation.

Fig. 9 – The data scientists toolkit to examine bias

Dataset profiling
User exploration into how protected and non-protected features are distributed within the dataset. Absolute or normalized views. Relationships between features and ground truth.

Protected feature impact analysis
Visualize the impact that protected features have on the average applicant. Categorize the protected features based on their impact to both local and global populations.

Protected group risk assessment
Analyze protected groups and assign a risk score based on the level of discrimination. Classify certain protected groups as "safe" and prioritize which other groups to further investigate.

Non-protected feature risk assessment
Highlight any unintended replication of inequalities from causal acting as proxies for protected features.

Model assessment and tracking
Assess fairness and performance per model iteration. Monitor the effects of different datasets, algorithms, tuning, and any combination thereof using configurable fairness and predictive performance metrics.

Conclusion

We have examined bias, what it is, why it is important, how it can creep into models, and how to manage associated risk. We conclude the paper by summarizing the six lessons we have learned:

Bias is tricky

The problem with bias begins with the layman's fallacy that it is well understood and can thereby be relatively easily avoided. The subtle truth paints a different picture. Keenly aware of this, academics have postulated many mathematical definitions of bias, often contradictory, each aiming to achieve a fair outcome. This holds true for each AI model: bias must be evaluated in ways compatible with the particular case in order to ensure the appropriate fairness objective.

Bias is complex

Unfortunately, bias is not simply a matter of black and white. Rather, it is actually a statistical result, which relies on observing different outcomes for populations who exhibit a particular protected feature such as a particular race, gender, religion, age or ethnicity. As we all know, these are not mutually exclusive features, so the statistical distributions themselves overlap. Bias is also practically unavoidable. While we will never fully stamp out bias, we should strive to treat people as fairly as possible. That means implementing decision models that balance appropriately between optimization objectives and fairness to the individual.

No silver bullets

How can a model be biased against protected classes, if never trained on that data to begin with? It feels almost too easy – and it is: removing protected features from a model will in most cases not remove bias. The reason is that seemingly neutral variables contain clues that lead back to the protected class. They are proxies for the very features we intentionally removed. Despite best intentions in removing gender as a factor, a model continues to discriminate on "assumed gender", having derived it from proxy features, such as occupation. The model learned from its training data that nurses are mostly female and construction workers mostly male.

More than performance

The good news is models can be designed to be less biased. The bad news: often this will come at a cost of lower discriminatory power – simply stated, the ability to decide between "yes" and "no". Optimizing on both performance and fairness is a complex task requiring skills and tools do execute properly. It is also not enough to get things right at launch. Populations change, the data changes, and model efficacy degrades over time – both in terms of performance and fairness. Modelers must be aware of this, both for static models in need of an update, as well as reinforcement learning models that update themselves. In both cases, a watchful eye is needed to ensure awareness for changing mix, changing proxies over time.

Beware of inept implementation

Machine learning models can also be made to learn bias – which is why legislators and regulators are increasingly proliferate in guidance, frameworks and rules on the subject. AI is already delivering tremendous commercial and scientific value – yet organizations are increasingly aware that they must be used with caution. Executives recognise that inexperience in managing the technology can yield unexpected, even harmful results. Worse, unlike people driven processes, they can do so systemically and in ways difficult to detect. Legal, regulatory and reputational exposure is disproportionately high compared to traditional, less inherently scalable approaches.

Values, skills and governance pave the way

Understanding the complex nuances of bias is the first major step towards effectively managing the risk. A suitable operating model and appropriately skilled staff are the necessary measures to act on this newfound awareness for this fairness dimension of model quality. Properly sensitized for bias, much of the existing modelling or software development operating model will suffice. That, together with correspondingly training for modelers (both to be conscious of bias and how to analyze it) will likely suffice to ward off the greatest dangers and associated unpleasant headlines – or fines.

Contacts

David Thogmartin
Director
aiStudio | AI & Data Analytics
dthogmartin@deloitte.de

Andy Whitton
Partner in Risk Advisory
Regulatory & Legal Support, Financial Service
awhitton@deloitte.co.uk

Michelle Lee
Senior Manager
AI Ethics Lead
michellealee@deloitte.co.uk

www.deloitte.com/de/aistudio

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms, and their related entities (collectively, the "Deloitte organization"). DTTL (also referred to as "Deloitte Global") and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/de/UeberUns to learn more.

Deloitte provides industry-leading audit and assurance, tax and legal, consulting, financial advisory, and risk advisory services to nearly 90% of the Fortune Global 500® and thousands of private companies. Legal advisory services in Germany are provided by Deloitte Legal. Our professionals deliver measurable and lasting results that help reinforce public trust in capital markets, enable clients to transform and thrive, and lead the way toward a stronger economy, a more equitable society and a sustainable world. Building on its 175-plus year history, Deloitte spans more than 150 countries and territories. Learn how Deloitte's more than 345,000 people worldwide make an impact that matters at www.deloitte.com/de.

This communication contains general information only, and none of Deloitte GmbH Wirtschaftsprüfungsgesellschaft or Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms or their related entities (collectively, the "Deloitte organization") is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related entities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication. DTTL and each of its member firms, and their related entities, are legally separate and independent entities.

Issue 01/2022