Proceedings of the 55th Hawaii International Conference on System Sciences | 2022

An Exploratory Study on Fairness-Aware Design Decision-Making

Sumaiya Sultana Tanu
University of Arkansas
sstanu@uark.edu

Lu Zhang
University of Arkansas
lz006@uark.edu

Dinesh Gauri
University of Arkansas
dgauri@walton.uark.edu

Zhenghui Sha
University of Texas at Austin
zsha@austin.utexas.edu

Abstract

With advances in machine learning (ML) and big data analytics, data-driven predictive models play an essential role in supporting a wide range of simple and complex decision-making processes. However, historical data embedded with unfairness may unintentionally reinforce discrimination towards minority groups when using data-driven decision-support technologies. In this paper, we quantify unfairness and analyze its impact in the context of data-driven engineering design using the Adult Income dataset. First, we introduce a fairness-aware design concept. Subsequently, we introduce standard definitions and statistical measures of fairness to the engineering design research. Then, we use the outcomes from two supervised ML models, Logistic Regression and CatBoost classifiers, to conduct the Disparate Impact and fair-test analyses to quantify any unfairness present in the data and decision outcomes. Based on the results, we highlight the importance of considering fairness in product design and marketing, and the consequences, if there is a loss of fairness.

1. Introduction

Decision-making based on big data is becoming increasingly popular within organizations with the growth of digitalization and remarkable advances in machine learning (ML) and artificial intelligence (AI). Online retail, car and home rental agencies, and airline industries, use algorithms to pool target audiences, determine customer perceptions, pricing strategies, and competing offers [1]. As the use of algorithms becomes prevalent, possible risks may develop and could give rise to discrimination. One typical example is the recommender system that studies user behavior and generates recommendations for users to support decision-making [2]. It can easily inherit bias in these datasets since most industries have been off-limits to protected classes (e.g., Women, Non-Caucasians, Senior individuals) [3]. Biased data can condition decision-making models to make unfair predictions without any deliberate intent of the developer or designer. Therefore, a design project's constraints nowadays move from strictly technical and economical to social, environmental, and ethical dimensions, mainly to incorporate the fairness consideration in design decision-making.

In the engineering design field, the investigation of fairness-aware algorithms is little observed. The lack of acknowledgment of discrimination perpetuated across design can reduce the access of unprivileged individuals to everyday tools and technology, widening the gap between unprivileged and privileged groups [4]. This could negatively impact a business, especially with the advent of social media and information being shared at lightning speed. For example, in Figure 1, we show human biases embedded in various stages of the standard engineering design process in developing and improving functional products and processes. In a data-driven decision support system, these biased human judgments and data could be translated into a product's design life cycle and various marketing strategies such as online targeted advertising. Potential harms can arise if ads are manipulative and stereotyping by targeting specific people and groups [5]. This can provoke unexpected customer behaviors in social media and can lead to many unintended consequences for businesses [6]. Hence, it is necessary to take fairness into account when products are designed and marketed.

To address the challenges of discrimination in these existing structures and dynamics of the societies due to their demographics, we study the fairness definitions and different statistical metrics quantifying unfairness. Barocas & Selbst [7] points out the relevance of the statistical data measures to the decision at hand to study fairness in data-driven approaches. This method can help investigate causes of discrimination in the datasets even when the goal is to ensure the greatest possible accuracy for its purposes. It can facilitate in exposing the exact magnitude of inequality in data. Using the metrics, we evaluate the outcomes from data-driven decision-support technologies and gain insights into the dataset analyzed. In this study, developing a fairness and social awareness platform in engineering design is twofold. First, an organization must put solid values,

URI: https://hdl.handle.net/10125/80013
978-0-9981331-5-7

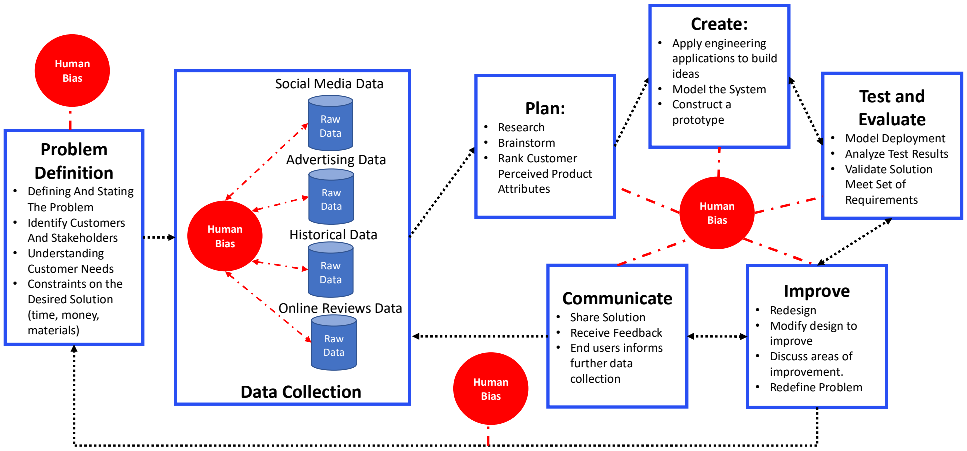


Figure 1. Biases in engineering design decision-making process

principles, and protocols to design and implement ethically and safely. Second, to facilitate a culture of responsible design innovation, various departments can work collaboratively to make fair design decisions and develop products, services, and technologies for the public [8]. A complex system such as engineering design can benefit from a fair and ethical strategy to better answer the following research questions: Is the dataset used insufficiently representative of the population? What are the top attributes that correlate with the decision of the ML models? How are design decisions impacted by drawing inferences from possibilities of biased and discriminatory outcomes? Lastly, how does the application of fairness statistical metrics aid with design guidance?

The remainder of the paper is structured as follows: Section 2 provides the relevant design concepts and problems that can be addressed by fairness consideration in design decision-making. Section 3 introduces standard statistical measures of fairness and the methodology to the engineering design research community. The proposed research approach is presented in Section 4, and the case study based on the Adult Income dataset is presented in Section 5. Section 6 discusses the results from the ML models and their impact on engineering design decisions. Finally, Section 7 concludes the paper with the limitations of the current research and future research opportunities in design for market systems with fairness.

2. Integrate Fairness in Engineering Design

2.1. Application of the Fairness Concept

Fairness is a subjective perception of situations, actions, or outcomes as being fair or unfair. Fairness is achieved when the ML model outcomes behave similarly for two or more classes of a group (e.g., male and female groups, majority and minority groups) [9]. Incorporating the concept of fairness can strongly influence the possibility of generating innovations or design expansion in engineering design practices [10]. Furthermore, it can improve the design in the progress of acquiring scientific-technological knowledge and the conditions of non-technical socio-economic factors. From a socio-economic perspective, it could provide individual customers with more exposure to products and services in their geographical context. Therefore, a fairness-aware design should be put in place to cope with the ethical dimension in the engineering design community.

2.2. Relevant Design Concepts for Fairness

In this section, we discuss several design concepts relevant to fairness-aware design applications.

2.2.1. Inclusive Design. Coleman states, "Inclusive design is not a new type of design but an intentional project that sets out to include significant sectors of society that are all too frequently ignored or overlooked." This concept incorporates diversity in design to address the needs of the broadest possible audience but does not intend to create a new genre of design that jeopardizes individual sectors of a society [11].

2.2.2. Design for Market Segments. Diverse customers will have different expectations. The market is segmented to divide a broad customer base into sub-groups of existing and prospective customers. By dividing the markets into smaller segments, we target customers that share characteristics such as everyday needs, shared interests, similar lifestyles, or even similar demographic profiles. It allows companies and organizations to develop products, services, and promotional campaigns targeting specific segments

[12]. Please note that fairness-aware design can be assured in each market segment.

2.2.3. Customized Design. A customized design gives users more control over their interactions with the system, such as customers can specify their individual needs and preferences, designs, and layout that appeal to them [13].

2.2.4. Personalized Design. Personalization is about building meaningful one-to-one relationships by understanding and meeting the needs of each customer, aiming at providing the customer with tailored products, services, information, or information related to products or services [14].

2.2.5. Fairness-aware Design. In this paper, we introduce the fairness-aware design concept. It is an extension of inclusive design in which a user's sensitive attributes do not play a significant role in design decisions. Constraints are applicable based on the product or service market segment; however, designs discard any discriminatory factors.

Based on the defining features for these design concepts, we present a few design scenarios and categorize them under single or multiple design concepts. It is worth noting that these design concepts are not mutually exclusive and may overlap in one particular design scenario.

Scenario 1: *Develop a video game that fits both male and female players.*

- Design case: Inclusive design and design for market segments. The market is segmented to understand the needs of male and female players in the market to cater to an inclusive design.

Scenario 2: *Design a recliner chair in movie theatres for people with and without disabilities.*

- Design case: Inclusive design, design for market segments, and customized design. This design falls into three design categories. From this scenario, it is understood there is a need for two variations of the recliner chair. Each variation is customized to best fit the market segment, i.e., disabled versus non-disabled individuals who enjoy going to movie theatres.

Scenario 3: *Recommendation on advertisements (movies, games, online retail) based on customer data and predictive technology.*

- Design case: Personalized design and design for market segments as information is tailored to a user's specific needs and preferences based on their online search patterns (personalized) and users with similar online behaviors (segmented).

Scenario 4: *Design a luxury car for all high-income group regardless of customer demographic*

- Design case: Fairness-aware design, design for market segments, and customized design. The intent is to target all eligible customers who can afford the price of the car, thereby maximizing the value of the luxury car. In this case, the luxury car is designed and marketed for members of the high-income group, regardless of their age, gender, race, or marital status. This market segment can also suggest individuals are able to customize their luxury cars.

The ethics dimension of fairness has emerged mainly due to the range of individual and societal harms that the misuse, abuse, poor design, or unintended negative consequences of engineering design systems may cause. To orient the reader to the concepts under discussion and the importance of building a robust culture of a fairness platform, we represent two of the most consequential forms that the potential unfairness may take in the following subsection.

2.3. Addressing Design Problems with Fairness Consideration

Problem 1: Bias in Machine Learning Decision Support Systems. As mentioned earlier, predictive models learned from historical data are widely used to help organizations make decisions. However, such decisions may mistreat individuals based on their respective attributes, increasingly raising concern about fairness and discrimination. For example, if a company decides to use historical data to extend its reach to new and potential customers, predictions made about them may be embedded with biases and discrimination against specific sectors of a society. When businesses tend to depend on such biased data, it causes unprivileged sectors of that society to be excluded from the market as target audiences. When studying the market, we study user's preferences and interactions with other users to improve and refine the products and services or develop new generations of products. However, depending on biased information, the system unintentionally continues to cater only to privileged society members, without knowing how the unprivileged groups might be affected. Application of fairness evaluation methods can help understand the attributes that affect eligible individuals by ML predictions and its further impact on engineering design decisions.

Problem 2: Bias in Marketing Strategies. Organizations study the market segments and their social network to analyze various users' product co-consideration and develop marketing strategies and advertising campaigns to target the correct audience. However, suppose the information used has a bias. In that case, a business may generate promotional campaigns that may contain discriminative content or exclude a sector of the population deserving a product or service. For example, multiple cases have noticed Facebook ads being used to exclude certain races, languages, and religious affiliations from advertisements for housing, credit, and insurance; and disproportionately target men over women for highly paid jobs. In light of these situations, studying the fairness of a marketing strategy will allow us to understand how negative information or content of an advertising campaign diffuses on social media platforms and the impact on the demand and profit of an organization.

3. Technical Background

Several statistical measures have been proposed in the ML literature [15-16] to define fairness such that if there exists a disparate distribution of a sensitive attribute, statistical fairness analyses are likely to depict that distribution and quantify the bias. This section explores the statistical metrics and various fairness definitions studied by fair ML scholars.

3.1. Statistical measures of fairness

Fair machine learning is a subset of machine learning and AI ethics that has gained attention in response to the rapid integration of machine learning into social realms [17]. An increasing number of decisions made by engineers regarding product design and its attributes are with the support of ML [18]. ML will facilitate interpreting big data analytics into visualization models, finding patterns, gaining insights in data, and using them to make predictions and expand their opportunities [19]. Integrating fairness-aware ML models in the design process will aid in identifying factors that are likely to compromise the enterprise and the customers. It builds trust, widens reach, and demonstrates to customers that their concerns matter.

A fair machine learning process is chosen to support design decision-making because the mathematical framework considers sensitive attributes for which non-discrimination can be established. We explicitly include gender and ethnicity attributes since such sensitive membership influences various stereotypes in engineering fields. The study is conducted in a binary classification setting, and throughout the paper, the terms protected or sensitive attributes are used indistinctly. The following notations will be used in the rest of the paper to understand the statistical measure of fairness:

- X: All qualified features that characterize an individual. It is represented in binary, numerically, or categorically in a dataset (e.g., location, age, demographics characteristics, loan repayment rates)
- A: Binary (0 or 1) sensitive or protected attributes of an individual. (e.g., race, gender, socio-economic status)
- Y: Target variable in the dataset that provides the actual classification result.
- S is the score that is predicted by the selected classifier is represented by $S = s(x, a) \in [0,1]$.
- $C = c(X, A) \in \{0,1\}$: Binary predictor (e.g., being shown the ad or not), which makes decisions based on a score S. For instance, if S is above a certain threshold (e.g., $s \geq 0.5$) then predicted outcome $c = 1$ is classified positive by a machine learning classifier and generally, it the preferred decision.

3.2. Evaluation of fairness in the U.S. legislation

There are two forms of discrimination generated in legal domains (1) disparate treatment and (2) disparate impact (DI). Disparate treatment is direct discrimination and an intentional act on individuals due to their sensitive attributes, such as age, gender, and race. DI is indirect discrimination that is unintentional by an enterprise's action yet disproportionately impacts individuals of a protected class due to bias in historical data [20-21]. Such as event is caused by redundant encoding that provides knowledge regarding a protected or sensitive membership based on features present in datasets that correlate with these memberships. For example, the purchase of video games and action movies online may be highly correlated with gender, and specific zip codes may have different racial demographics that an ML algorithm has learned from a classifier. While such algorithms are unlikely to generate disparate treatment, data-driven algorithms trained with such biased datasets are likely to grasp these biased patterns unrecognized by humans and cause disparate impact [22].

A DI index value is used in regulated domains to evaluate whether a decision-making system is free of disparate treatment and disparate impact. In fair machine learning literature, it is mathematically represented by Equation (1) [19, 21] to measure unfairness that exists in datasets and is used to quantify a group fairness [23].

$$DI = \frac{P(Y = 1 | A = unpriviliged)}{P(Y = 1 | A = privileged\,)} \qquad (1)$$

3.3. Definitions of Fairness

Based on the notions and statistical metrics of the confusion matrix, the fairness functions can be formulated. These definitions are centered on the predicted probability, predicted outcome, actual outcome (target variable), and correlation. There are various definitions of fairness proposed in algorithmic fairness literature, most of which are derived from the following five metrics:

1) Group Fairness

This notion studies the impact of discrimination on a group of individuals and therefore uses ML classifiers to mitigate unfairness in models. This definition is satisfied when a supervised ML classifier predicts the same probability outcome for individuals in both the privileged (e.g., males = m) and unprivileged (e.., female = f) categories. This probability will suggest that both parties are receiving equal predicted positive ($C = 1$) outcomes; hence the model is fair. This category branches three fairness definitions.

a. Demographic Parity

This states that the binary predictor C is independent of A and the proportion of each segment of a protected class (e.g., gender) should receive the positive (or negative) classifications at equal rates. In this case, the probability measure [2] is formalized as,

$$P(C = 1|A = m) = P(C = 1|A = f) \quad (2)$$

b. Equalized Odds

This fairness notion is independent of A but is conditional on actual outcome Y such that the probability for positive prediction is equal for both parties regardless of Y outcome. Therefore, both groups have equal recalls and are satisfied by the classifier [24].

$$P(c = 1|Y = 1, A = m) \\ = P(c = 1|Y = 1, A = f) \quad (3i)$$

$$P(c = 1|Y = 0, A = m) \\ = P(c = 1|Y = 0, A = f) \quad (3ii)$$

c. Predictive Rate Parity

This is based on both the predicted and actual outcomes. Here the probability of the positive predicted outcome for both males and females to be also the actual outcome. This probability states that to be fair, the correct positive predictions are the same for both parties in equal fractions. This is also true for the negative predicted outcome from a mathematical standpoint [25].

$$P(Y = 1|c = 1, A = m) \\ = P(Y = 1|c = 1, A = f) \quad (4i)$$

$$P(Y = 0|c = 1, A = m) \\ = P(Y = 0|c = 1, A = f) \quad (4ii)$$

2) Unawareness

Fairness through unawareness is simply when no sensitive attributes, A, are used to train the classifier, so the predicted and actual outcomes are not influenced by A. Here, the predicted outcome of individuals with the task-specific features, X, is the same as shown in Equation (5). The probability is [26]

$$C = c(X_i) = c(X_j) \quad (5)$$

3) Individual Fairness

As the famous quote Dwork formalizes, ML models should treat similar individuals similarly. A distance metric defines the similarity for the fairness to hold. In our study on the fairness-aware design, this notion is used to study customer-related attributes based on the individual level (local level) information obtained from group analysis [15].

4. Research Approach Overview

In this study, we begin with tackling Problem 1 (see Section 2.3) as a case study. We apply the fairness evaluation methods familiarized in Section 3 and quantify any bias present in the decisions made by ML models in the context of engineering design.

The first step of the study is to conduct a thorough pre-processing of raw data. Data-preprocessing is a critical step before fitting into a statistical model or training a classifier on the dataset. We remove all missing information from the data to minimize discrepancies. In the dataset, each row corresponds to an individual's attribute. Attributes that do not have a clear meaning regarding the analysis or the individual are removed from the data. Data with large differences were merged to eliminate redundancies. Categorical attributes were trimmed to remove space from a cell.

After the data pre-processing, the data is prepared for supervised machine learning models. Dataset is split into 80% training data, 10% verification data, and 10% testing data following the standard ten-fold cross-validation procedure. The k-fold cross-validation randomly splits the dataset into k equal-sized data sets. It uses k-1 sets as a training dataset while using the remaining one dataset as a test dataset. The model repeats the task until each dataset is used for testing. We use two supervised ML models: Logistic Regression (L.R.) and CatBoost (C.B.), to predict binary outcomes for the test data. Generally, a binary classification outcome involves two class labels: the normal/majority state (labeled 1) and the abnormal/minority state (labeled 0). Binary classification helps us isolate vast quantities of data into discrete values such as 0/1, True/False, or a pre-defined output label class. For example, predicting the admission of a potential university applicant using ML predictive models, "accepted" is the normal state, and "not accepted" is the abnormal state.

In the first step, we analyze the correlation of attributes concerning the target variable and compare the prediction accuracies to understand the performance of each classifier. Next, we test for disparate impact calculated using the predicted probability and the binary outcomes in the second step. Classifiers L.R. and C.B. are then trained on two sets of data (1) dataset with sensitive attributes (2) dataset with sensitive attributes removed. In the third step, we use predicted probability, s, to test fairness based on the attributes with the most severe disparate impact. In both analyses, we use conditional probability, Equations (1), (6), and (7) to verify that the fairness is satisfied or worsened for groups $A = 1$ and $A = 0$. Based on the results, the final step consists of studying the impact of ML predictions on engineering design decisions to guide the design towards fairness.

5. Case Study

This case study investigates the effects of individual characteristics, standard attributes (e.g., education level, occupation, country), and sensitive attributes, such as gender, race, age, on the prediction of the target variable, income status. The dataset is the open-source Adult Income data accessible through the University of California Irvine (UCI) Machine Learning repository. Extraction of this dataset was conducted by Barry Becker from the 1994 Census database [27].

The dataset consists of 48842 individual records with six continuous attributes (age, final weight, education level, capital gain, capital loss, and hours per week) and eight categorical attributes (work class, education, marital status, occupation, relationship status, race, sex, country). Income attribute is the target variable, i.e., a binary variable indicating whether an individual's annual income is greater than $50,000 a year or not, i.e., $Y = 1$ if the income is higher than $50,000 and $Y = 0$ if the income is less than or equal to $50,000. Learning about customer requirements and the disposable income they can spend is crucial in developing a new product. The outcomes from this step strongly influence the rest of the development effort and the ultimate success of the product.

After data pre-processing, the dataset consists of 30162 applicants. Education and Education Level are merged in the analysis as it represents the same task-specific attribute which trimmed down the dataset to nine attributes. All attributes are changed to either binary or continuous variables. Figure 2 provides a graphical representation of the distribution of the important attributes and shows the count of occurrences of each attribute with respect to the actual income classification.

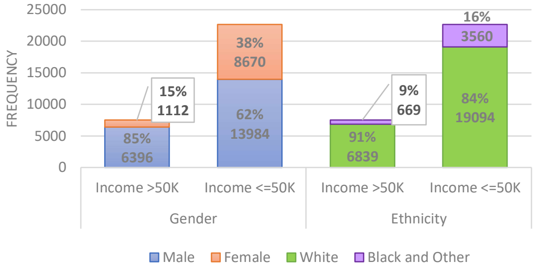


Figure 2. Sensitive attributes of the Adult Income dataset after data pre-processing

Based on the preliminary exploratory data analysis, we observe an unbalanced repartition of low and high incomes concerning two variables: Gender (male or female) and Ethnic origin (White, Black, and Other). It is observed that the sample contains only 15% female and 9% non-Caucasians (Black and other combined) who earn an income higher than $50,00 compare to male and White individuals, respectively. Therefore, we categorize these attributes as sensitive attributes. In the next section, we first identify the correlation between each of the attributes and decisions made by the ML models. Second, we use two fairness evaluation methods: disparate impact and fair-test analysis, to evaluate the data-driven decision and quantify any potential biases in the dataset. Lastly, we discuss the impact of the potential biases and discriminatory outcomes on design decisions.

6. Results and Discussion

6.1. Machine Learning Algorithm

In this study, two ML classifiers were used to train the Adult income dataset and make predictions for the test data.

Logistic Regression (L.R.) Classifier: In the L.R. model, the response has two possible outcomes ($C = 0$ or $C = 1$). A heat map grid shown in Figure 3 is used to visualize the correlation matrix of the attributes before modeling the dataset. A higher positive value suggests a strong correlation between features and target outcome as well as a higher chance of target outcome to receive $C = 1$ ($50,000 and above). The correlation coefficients are generated using the Pearson method. This method is used to determine if a significant linear relationship (positive, negative, or zero) exists when

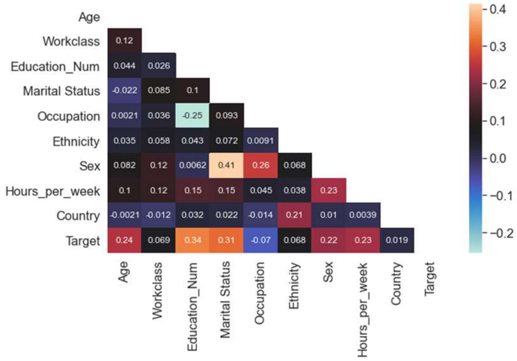


Figure 3. Heat map showing the correlation between attributes in adult income dataset

two quantitative variables (e.g., the target variable vs. the sensitive attributes) are being tested in this study. In this matrix, the sensitive attributes (age and gender) strongly correlate with the target outcome, with a value of 0.24 and 0.22, while ethnicity has a weaker correlation with a value of 0.068. However, based on Figure 2, it is evident that there is an unbalanced division between the income prediction for the ethnicity and gender attributes. Therefore, based on the correlation of sensitive features with the target variable and the distribution of income, we have grouped male (m) and White (w) as privileged groups ($A = 1$) and grouped protected membership, such as women (w), black and other minorities as an unprivileged group ($A = 0$).

CatBoost Classifier: This classifier is based on gradient boosting on decision trees. It generates SHAP values, representing the importance of the dependent variables, and measures the impact of attributes with respect to the target variable [28]. This classifier is trained on the Adult Income dataset, and the SHAP summary is plotted in Figure 4. This figure is interpreted at an aggregated level since the summary looks at the entire dataset.

Each point in this plot corresponds to one observation from the dataset. The x-axis is the SHAP value that quantifies the probability of success for that attribute i.e., a high positive SHAP value indicates that this feature will drive the prediction of an individual towards $C = 1$ (the income is more than $50,000), and a negative SHAP value indicates the feature will drive a prediction towards low income, $C = 0$ (i.e., the income is less than and equal to $50,000). The attributes are ranked based on the order of importance. That means the attributes having the strongest correlation with the target variable and most influential to the prediction will be ranked at the top of the SHAP model. The y-axis represents the attribute name. The color gradient specifies the numerical or categorical value of an attribute [28].

6.2. Fairness Measures with Disparate Impact

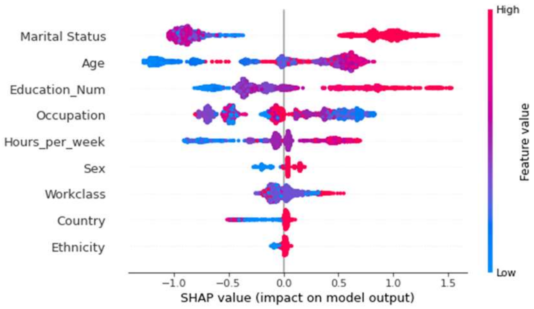


Figure 4. SHAP values of the CatBoost classifier

Table 1. Disparate impact using actual outcomes (the reference scores)

	With Sensitive Attributes
Gender	0.36
Ethnicity	0.59

In order for a decision-making system to be fair, the disparate impact index should be close to 1. A lower index value indicates that the discrimination is stronger over the unprivileged group. In this study, we first calculate the DI index based on the actual outcome using Equation (1), which serves as the reference point, DI_{ref}. The index value for gender is 0.36, and ethnicity is 0.59, as shown in Table 1. The classifiers L.R. and C.B. are then trained on two sets of data: (1) Adult income data with sensitive attributes and (2) Adult income data without two sensitive attributes (gender and ethnicity removed). In this approach, we calculate the DI index values based on the predicted outcome C using Equation (6) [29]. This breaks down this study into four categories:

- $i = 1$: Dataset trained without sensitive attributes using L.R. classifier
- $i = 2$: Dataset trained without sensitive attributes using C.B. classifier
- $i = 3$: Dataset trained with sensitive attributes using L.R. classifier
- $i = 4$: Dataset trained with sensitive attributes using C.B. classifier

This approach is taken to see if there will be an increase in DI_1 and DI_2 values for dataset without the sensitive attributes in comparison to DI_3 and DI_4**.** Since it is assumed that there are no sensitive attributes, there is no discrimination among groups $A = 1$ (gender is male and ethnicity is white) and $A = 0$ (gender is female, and ethnicity is black and other combined). The results are shown in Table 2.

$$DI_i = \frac{P(C = 1|A = 0)}{P(C = 1|A = 1\,)} \qquad (6)$$

The DI_{ref} calculation was compared to the predictions made by the respective classifiers DI_i. Based on the results observed in Table 2, it is identified that the disparate impact for gender and ethnicity for DI_1 worsened with the removal of sensitive attributes using L.R. However, for DI_2, the C.B. was not impacted by the removal of sensitives attributes. Standard regulations promote the removal of sensitive attributes when dealing with the fairness of machine learning algorithms. However, in this analysis, the removal seems irrelevant as removing the sensitive attributes solely is not sufficient to make a model fair. As for DI_3 and DI_4, both the classification algorithms C.B. and L.R. performed similarly with the gender attribute having more bias than ethnicity. We observe that data-driven decision-support systems have replicated,

Table 2. Disparate impact using predicted outcomes

Category *i*	$i = 1$	$i = 2$	$i = 3$	$i = 4$
Gender	0.25	0.34	0.23	0.26
Ethnicity	0.19	0.56	0.47	0.43

reinforced, and amplified the patterns of inequality and discrimination in the new set of data, i.e., test data. Likewise, because many of the features, metrics, and logical structures of the models enable the designers to choose data mining, these technologies can potentially replicate their designers' preconceptions and biases. Therefore, concluding such predictions can create real possibilities of discriminatory outcomes because the data fed into the systems is flawed from the start [8].

6.3. Test Fairness of Calibration Scores

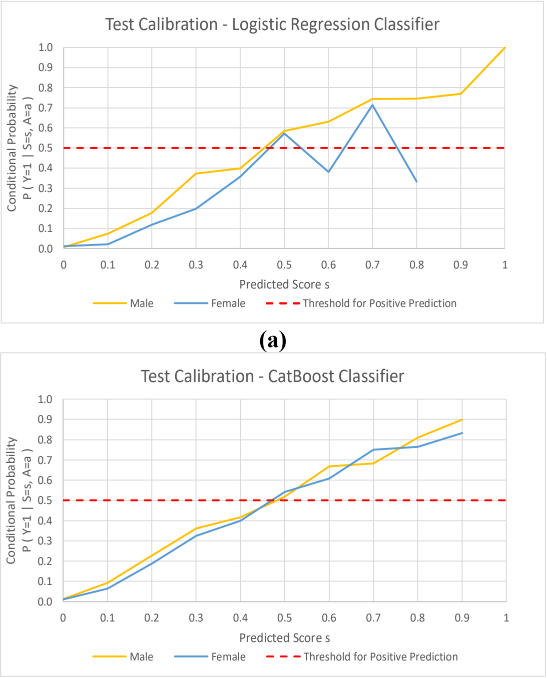
This analysis quantifies fairness using calibration scores based on predicted probability values and the actual outcome. The calibration score is calculated using Equation (7) [16]. In the previous sections, we see that the gender attribute on average has a severe DI compared to the ethnicity attribute and has the strongest correlation to the target variable. Hence, we focus on the gender attribute to determine if the calibration score explains an individual with an actual positive outcome $Y = 1$, but to have a predicted outcome of $C = 0$. A calibration score is test-fair (well-calibrated) when both males and females have the same probability. The predicted probability of each individual in the test dataset is calculated and shown in Figure 5. The threshold for the calibration score is set to 0.5. This is seen as the red dashed line on both the graph in Figures 5(a) and 5(b). Beyond this threshold score, the predicted probability is likely to return a positive binary prediction of $C = 1$ for individuals in test data. This means that an individual who received a score greater than 0.5 will have the prediction of $C = 1$ as their actual outcome $Y = 1$. Calibration scores for both privileged and non-privileged members should be equal to satisfy this fairness metric given by Equation (7).

$$P(Y = 1|S = s, A = m) = P(Y = 1|S = s, A = f) \qquad (7)$$

In Figure 5, the candidates on test data were randomly selected by the classifiers. From the first observation in Figures 5(a) and 5(b), it indicates the C.B. classifier satisfies the Equation (7) for higher values of predicted score s (greater than 0.5) than L.R. However, for both classifiers, the predicted scores s in the lower range between 0 to 0.5, Equation (7) is not satisfied. This is because both males and females are likely to have an unfavored prediction (i.e., $C = 0$) even if the actual outcome is $Y = 1$. For example, a male with an actual income greater than $50K is predicted to have a lower income by the classifiers. Second, for a probability score greater than 0.5, L.R. does not perform well and doesn't satisfy the notion for the sensitive attribute, female, even for the high probability scores. This is observed in Figure 5(a) when there is a sudden drop in the conditional probability (y-axis) for sensitive attribute $A = f$ and the prediction score s values of 0.6 and 0.8 (high predicted probability). In this case, a female candidate with the same s scores did not receive the same treatment as a male applicant even when their actual outcome is $Y = 1$.

6.4. Impact of ML Predictions on Design Decisions

Fairness statistical analysis indicates that eligible candidates ($Y = 1$) who received a prediction of $C = 0$ cannot afford a product or a service a business offers. The decision recommended by the ML model will automatically reject them as the target audience, and customers that participate in the design process (e.g., focus groups) are often members of the privileged group. Therefore, using the ML decision support system unintentionally leaves a trail of bias throughout the entire design decision-making. With design methodologies for market systems being popular, this may result in designing a product that only caters to a specific sector of the eligible society. For example,



(a)

(b)

Figure 5. Fair-test analysis using predicted probability scores

suppose we exclude non-Caucasian female participation and point of view from developing a video game design since the prediction received is Y=0. In that case, they either (a) create a female role as background characters and not a leading player or (b) design a game and marketing strategies exclusive to male players. In this scenario, the organization is inappropriate hurting. It has a disparate impact on non-Caucasian female individuals when they are eligible members in the unprivileged group, i.e., they earn more than $50,000. Since machine learning algorithms are to minimize the loss function, the algorithms will naturally favor groups contributing more to the training process (i.e., the majority groups) and less favorable to minority groups. Such biases could degrade population retention in minority groups and exacerbate representation disparity. Without incorporating the ethical dimension of fairness in the decision process, the minority groups are likely to diminish, and the vendor will lose the market on these groups [30].

Fairness evaluation methods can ensure bias is quantified in the datasets and capture as many eligible and potential individuals as possible with the highest accuracy possible. Incorporating customer preferences and opinions from privileged and unprivileged groups as input into the product design processes can result in more effective and fair design decision-making. The fairness-aware design approach will maximize the product's value in terms of social, humanistic, and economic values. A positive diffusion social network will propagate and drive more business among customers, converting into market demand. This also opens avenues in collecting more data from various sectors of the population and, in the future, has minimal chances of getting rejected as the target audience. It will generate business opportunities to create a diverse social network and develop fairness-aware marketing strategies.

6.4.1. Trade-offs. It is important to note, however, with fairness applications in design and data, a trade-off may exist between (1) impact on product demand and fairness (design) and (2) accuracy and fairness (data). This is, as we practice a higher degree of fairness, we may compromise on accuracy, demand, and design challenges. While there are studies conducted in literature for the trade-off between fairness and accuracy that allows for higher fairness without significantly compromising the accuracy or other concepts of utility [31], research explorations are limited or less investigated concerning trade-offs between product demand and fairness.

7. Conclusion and future directions

In this paper, we lay out design scenarios to understand the relevant design concepts of fairness. We then introduce the standard definitions and statistical measures of fairness from ML literature and conduct an exploratory study on the Adult Income dataset. The metrics disparate impact and fairness testing were analyzed to quantify any bias in the data between two classes of each sensitive attribute: privileged (majority) and unprivileged (minority) groups. Fairness is achieved when the ML model outcomes behave similarly for both groups, and sensitive attributes do not significantly affect the prediction. We first observe that the C.B. classifier model performs better than the L.R. classifier for each fairness evaluation method implemented from our analysis. Second, after the Disparate Impact analysis, we observed that gender attributes had a severe disparate impact value than ethnicity attributes in both training and testing data. This may cause discrimination as the Gender attribute can drive the ML prediction of an individual with $Y = 1$ to $C = 0$.

This research, in its current stage, is subject to a few limitations. First, although we can quantify fairness in datasets, direct application of the knowledge from computer science literature to design decision-making may be challenging due to the unique characteristics in the design and development process based on product characteristics. Second, if actual customer data was used to find potential customers, it may be challenging to quantify unfairness based on statistical measures that use actual outcomes. This would require a rigorous optimization of training data until an accuracy close to 100% is reached before analyzing new data.

The future scope of this work is twofold. First, we want to evaluate the design for market systems by tackling Problem 2: bias in marketing strategies. Our goal is to study the diffusion dynamics on a network of user-generated content of a product's advertising campaign on various social media platforms and its influence on market demand. Second, we will explore various data pre-processing approaches such as sampling or re-weighting the data to counterbalance discriminatory effects [32] and changing the individual data records [33] to help mitigate discriminatory bias in our problem in our future work.

8. Acknowledgement

The authors gratefully acknowledge the financial support from NSF through grant #1946391 and the Arkansas Economic Development Commission's Division of Science and Technology.

9. References

[1] I.Graef, “Algorithms and Fairness: What Role for Competition Law in Targeting Price Discrimination towards Ends Consumers”, The Columbia Journal of European Law, vol. 24, no. 3, 2018, pp. 541–559.

[2] S.Yao and B.Huang, “Beyond Parity: Fairness Objectives for Collaborative Filtering”, Conference on Neural Information Processing Systems, 2017, pp. 2.

[3] N.A.Saxena, et al., “How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness”, AAAI/ACM Conference on A.I., Ethics, and Society, 2019, pp. 99–106.

[4] G. Burleson, “Racism in America, Manifested in Engineering Design: A Pledge to Take Responsibility and Action”, Medium, The BYU Design Review, 2020.

[5] S.Barocas, M.Hardt, and A.Narayanan, “Fairness and Machine Learning: Limitations and Opportunities”, 2021, pp. 145-147.

[6] I. Horn, T. Taros, S. Dirkes, et al. “Business Reputation And Social Media: A Primer On Threats And Responses”, Journal Of Direct, Data And Digital Marketing Practice, vol. 16, 2015, pp. 193–208.

[7] S.Barocas, and AD Selbst, “Big Data’s Disparate Impact”, California Law Review, 2016, pp. 671–732.

[8] D.Leslie, “Understanding Artificial Intelligence Ethics And Safety: A Guide For The Responsible Design And Implementation Of A.I. Systems In The Public Sector”, The Alan Turing Institute, 2019.

[9] A.Chouldechova and A. Roth,” The Frontiers of Fairness in Machine Learning”, Recent Trends in Learning From Data. Studies in Computational Intelligence, vol. 896. Springer, Cham, 2020.

[10] H.E. Sladovich, “Engineering as a Social Enterprise”, National Academy of Engineering, The National Academies Press, Washington, DC, 1991.

[11] R. Luck, “Inclusive design and making in practice: Bringing Bodily Experience Into Closer Contact With Making”, Design Studies, vol. 54, 2018, pp. 96-119.

[12] M.A. Camilleri, “Market Segmentation, Targeting and Positioning”, Travel Marketing, Tourism Economics and the Airline Product, Springer International Publishing, Cham, 2017, pp. 69–83.

[13] J. Tiihonen and A. Felfernig, “An Introduction To Personalization And Mass Customization”, Journal of Intelligent Information Systems vol. 49, 2017, pp. 1–7.

[14] C. Adolphs and A. Winkelmann, “Personalization Research in E-Commerce - a State of the Art Review (2000-2008)”, Journal of Electronic Commerce Research, vol 11, no. 4, 2010, pp 326-341.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness Through Awareness”, Innovations in Theoretical Computer Science Conference, 2012.

[16] S. Verma and J. Rubin, “Fairness Definitions Explained”, Proceedings of the International Workshop on Software Fairness, 2018, pp. 1–7.

[17] T. Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines”, Minds and Machines, 2020.

[18] J.H. Panchal, et al. “Special Issue: Machine Learning for Engineering Design”, Journal of Mechanical Design, vol. 141, no. 11, 2019.

[19] S. Liu. “Towards Better Analysis of Machine Learning Models: A Visual Analytics Perspective”, Visual Informatics, vol. 1, no. 1, 2017, pp. 48–56.

[20] M.B. Zafar, et al. “Fairness Constraints: A Flexible Approach for Fair Classification”, Artificial Intelligence and Statistics (AISTATS), 2017.

[21] R.L. O’Brien, and B. Kiviat. “Disparate Impact? Race, Sex, and Credit Reports in Hiring”, Sociological Research for a Dynamic World, vol. 4, 2018.

[22] S. Matthew, “Programming Fairness in Algorithms”, Medium, Towards Data Science, 2020.

[23] D. Pessach et al., “Algorithmic fairness”, 2020.

[24] M. Hardt, et al., “Equality of Opportunity in Supervised Learning”, Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, 2016.

[25] A.Chouldechova, “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments”, Big Data, vol. 5, no. 2, 2017.

[26] J. Chen, et al. “Fairness Under Unawareness”, Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 339–348.

[27] K. Ron, “Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid”, Knowledge Discovery and Data Mining, 1996.

[28] L.S.Shapley, “A Value For N-Person Games”, Contributions to the Theory of Games, Princeton University Press, 1953, pp. 307-317.

[29] P. Besse et al., “A Survey of Bias in Machine Learning Through the Prism of Statistical Parity for The Adult Data Set”, 2020, pp. 1-25.

[30] X. Zhang, et al., “Group Retention When Using Machine Learning In Sequential Decision Making: The Interplay Between User Dynamics And Fairness”, In Advances in Neural Information Processing Systems, 2019.

[31] J. Kleinberg et al., “Inherent Trade-Offs in the Fair Determination of Risk Scores”, In 8th Innovations in Theoretical Computer Science Conference, 2017.

[32] F. Kamiran and T. Calders, “Data Pre-Processing Techniques For Classification Without Discrimination”, Knowledge and Information Systems, 2012, pp. 1–33.

[33] S. Hajian and J. Domingo-Ferrer, “A Methodology For Direct And Indirect Discrimination Prevention In Data Mining”, IEEE, 2013.

[34] R. K. E. Bellamy et al., “AI Fairness 360: An Extensible Toolkit for Detecting And Mitigating Algorithmic Bias”, IBM Journal of Research and Development, 2019.