Journal of Business Research 144 (2022) 93–106

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods

Benjamin van Giffen [a,*], Dennis Herhausen [b], Tobias Fahse [c]

[a] University of St.Gallen, Institute of Information Management, St. Gallen, Switzerland
[b] Vrije Universiteit Amsterdam, School of Business and Economics, Amsterdam, the Netherlands
[c] University of St.Gallen, Institute of Information Management, St. Gallen, Switzerland

ARTICLE INFO

Keywords:
Machine learning
Artificial intelligence
Bias
Mitigation methods
Case study

ABSTRACT

Over the last decade, the importance of machine learning increased dramatically in business and marketing. However, when machine learning is used for decision-making, bias rooted in unrepresentative datasets, inadequate models, weak algorithm designs, or human stereotypes can lead to low performance and unfair decisions, resulting in financial, social, and reputational losses. This paper offers a systematic, interdisciplinary literature review of machine learning biases as well as methods to avoid and mitigate these biases. We identified eight distinct machine learning biases, summarized these biases in the cross-industry standard process for data mining to account for all phases of machine learning projects, and outline twenty-four mitigation methods. We further contextualize these biases in a real-world case study and illustrate adequate mitigation strategies. These insights synthesize the literature on machine learning biases in a concise manner and point to the importance of human judgment for machine learning algorithms.

1. Introduction

Over the last decade, insights obtained from machine learning (ML) embedded in artificial intelligence (AI) revolutionized and fundamentally changed almost every aspect of daily life.[1] For example, ML algorithms make movie recommendations, suggest products to buy, decide on loan applications, and influence hiring decisions (Bogen & Rieke, 2018; Cohen et al., 2019). There are clear benefits when ML algorithms, in place of humans, make decisions: unlike humans, they are not susceptible to fatigue or boredom and they can take into account many more factors in their decision-making (Danziger et al., 2011). Not surprisingly, there is ample interest within the business and marketing domain to understand the opportunities presented by ML and AI (e.g., Davenport, Guha, Grewal, & Bressgott, 2020; De Bruyn, Viswanathan, Beh, Brock, & von Wangenheim, 2020; Ma & Sun, 2020; Wang, Ryoo, Bendle, & Kopalle, 2021).

However, like humans, ML algorithms are vulnerable to biases that make their predictions and decisions "unfair" (Angwin et al., 2016). In the context of ML decision-making, fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics (Mehrabi et al., 2019). Thus, a biased and unfair ML algorithm makes decisions that are skewed toward a particular group of people. Although ML algorithms operate in the digital domain, ML biases have many real-world consequences and may cause substantive harm to both consumers and companies. A famous example relates to the Apple credit card, launched in partnership by Apple and Goldman Sachs, which offered lower lines of credit to women than to men of equal or even lower financial standing (Vigdor, 2019).

Thus, business and marketing managers as well as researchers in these areas need insights regarding the biases that challenge the opportunities presented by ML. While some literature reviews of ML and AI in business and marketing already exist (e.g., Guha et al., 2021; Huang & Rust, 2021; Puntoni et al., 2021), all these studies only briefly touch on ML biases. This paper addresses this gap with an interdisciplinary literature review and an in-depth case study. Prior research has underlined ML's value for marketing automation, decision-making, as well as for analyzing interactions and human emotions in marketing research, strategy, and action (Huang & Rust, 2021). At the same time, a recent

* Corresponding author.
E-mail addresses: benjamin.vangiffen@unisg.ch (B. van Giffen), dennis.herhausen@vu.nl (D. Herhausen), tobias.fahse@unisg.ch (T. Fahse).

[1] The terms ML and AI, while so commonly used, remain poorly defined and fuzzy concepts (De Bruyn et al., 2020). In line with Ma and Sun (2020), for us AI is manifested by machines that exhibit aspects of human intelligence while ML refers to computer programs that are able to learn, adapt, and improve without following explicit instructions.

https://doi.org/10.1016/j.jbusres.2022.01.076
Received 18 January 2022; Accepted 23 January 2022

bibliographic analysis of AI in marketing, consumer research, and psychology identified ML bias as a highly relevant, but surprisingly neglected research topic (Mariani et al., 2021). Hence, scholars have urged for research that helps to "identify bias in relatively nascent AI applications, before much harm is caused" (Guha et al., 2021, p. 35). Against this background, our study seeks to shed light on ML bias in marketing.

First, the term bias is often used in a wide sense to refer to a variety of adverse effects caused by ML applications (e.g., unfairness or discrimination). We find many examples for ML bias, but little systematicity or theoretical guidance for framing and effectively analyzing bias in ML. To enable more effective discussion and potential mitigation of various types of ML bias, a comprehensive presentation and delineation of ML biases is warranted.

Second, the link between bias cause and effect is often obscure or not immediately obvious. Many infamous examples of ML bias point to flawed training data as a cause of bias (e.g., Weissman, 2018), even though "algorithmic biases arise from flawed generated processes […]" (Rai, 2020). As of now, there is no comprehensive framework for defining how different types of bias occur in the ML process and how these biases might be mitigated (or prevented).

Third, given the novelty of ML biases and its lack of prominence in marketing education, it is necessary to raise marketeers' awareness of potential biases and to build knowledge about ML to avoid biases (c.f., Huang & Rust, 2021; Puntoni et al., 2021). Hence, ML biases should be presented in a comprehensible manner so that marketing researchers and practitioners can effectively manage and address them in their ML projects.

We address these gaps by reviewing the largely disconnected literature on ML biases from different fields and by providing a shared terminology of mitigation methods to prevent these biases. Specifically, we use CRISP-DM, the widely adopted cross-industry standard process for data mining (Wirth & Hipp, 2000), as the underlying framework for our analysis and map eight different ML biases and twenty-four mitigation methods into the different process phases of an ML project. We then use a case study to illustrate the identified ML biases and mitigation methods in a marketing context. Thereby, we hope to inform and sensitize researchers and managers alike about the specific biases that might be present in their ML projects and the methods to mitigate their impact.

2. Conceptual background

2.1. Artificial intelligence and machine learning in marketing

The adoption of AI technologies across organizations is growing rapidly, and firms are increasingly recognizing the practical opportunities arising from their ability to perform human-like tasks such as learning autonomously or making decisions based on large datasets (Huang & Rust, 2021). In particular, AI has witnessed impressive breakthroughs in image recognition, speech processing, autonomous driving, and many other tasks typically considered to require human-level intelligence (Davenport et al., 2020). Behind much of this breakthrough is ML, which has become the main paradigm of contemporary AI research (Wang, Ryoo, Bendle, & Kopalle, 2021).

ML is a vast and rapidly evolving field, encompassing a wide range of methods for addressing diverse tasks (Ma & Sun, 2020). Despite this variety, the typical application logic of ML in marketing is summarized in Fig. 1. Data is generated from the relevant population which is used to train a predefined ML model that often classifies observations or optimizes a predefined outcome, and its predictions then trigger marketing decisions and actions. For example, Netflix generates data from the viewing behavior of all its customers, uses this data to train a recommendation algorithm, whose predictions then trigger individual movie and series recommendations for all its customers. Building upon this logic, a diverse set of ML methods, such as support-vector machines (Cui & Curry, 2005), topic models (Tirunillai & Tellis, 2014), ensemble trees

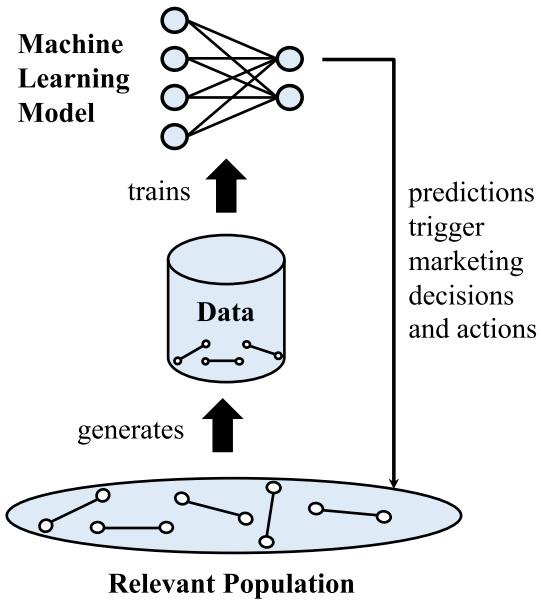

Fig. 1. Conceptual Approach: Using Machine Learning for Marketing Problems.

(Yoganarasimhan, 2020), and deep neural networks (Liu et al., 2020), have been used in marketing for making predictions that trigger decisions and actions.[2]

2.2. Reasons for machine learning biases

The decision-making of algorithms is very different to human decision-making in two important ways: algorithms are extremely literal and they are black boxes. First, while humans understand soft goals and trade-offs, algorithms will pursue a specified objective single-mindedly (Luca et al., 2016). For example, Ukanwa and Rust (2020) show that for loan decisions, discriminatory results can occur even if there is no bigotry programmed into the algorithm because the algorithm only seeks to maximize profit. Second, algorithms are black boxes in the sense that they can often form predictions with great accuracy, but they do not provide causes or reasons for an event. Thus, algorithms often lack interpretability, in terms of having a transparent model structure and clear linkage between variables (Ma & Sun, 2020).

Moreover, ML algorithms differ substantially from deterministic, rule-based algorithms that have been used in the past for decision support in the organizational context (Wang, Ryoo, Bendle, & Kopalle, 2021). ML algorithms, such as neural networks, follow a probabilistic approach in which decisions are not made by following programmed rules but by learning patterns from historical data and applying these to new input data. The decision support from ML algorithms is provided in the form of probabilities, leading to different levels of uncertainty and therefore increased susceptibility to systematic biases. For instance, Lambrecht and Tucker (2019) have shown that gender bias can occur without any conscious (or unconscious) attempt to produce a biased outcome—using only an unbiased algorithm.

Finally, a series of subjective choices must be made in the process of any ML project, and all these choices may introduce biases and lead to unwanted outcomes. For example, considering the logic of Fig. 1, not all relationships within the relevant population necessarily generate data, human coding might determine the data generation process, and the

[2] It must be noted that relationships uncovered using ML are often correlational rather than causal. With a predictive focus, little attention has been paid to endogeneity concerns. Issues such as selection, omitted variables, and simultaneity, which are addressed in econometric models, are typically ignored in ML algorithms.

impact of the ML model may reinforce certain patterns in the data generation. As a result, the data does not represent the "whole" relevant population because not all observations and relevant variables are recorded. Even if the data is perfectly unbiased, the decision on how to build and train the model can introduce biases (e.g., selection of unsuitable variables and over- or underfitting during the model training). Even if one assumes the resulting ML application is free from bias introduced through data or design decisions, an inappropriate context of use may nevertheless lead to a bias.

Research in business, marketing, computer science, psychology, and sociology has started to consider ML biases (e.g., De Bruyn et al., 2020; Guha et al., 2021; Huang & Rust, 2021; Puntoni et al., 2021). However, to date only a few scattered articles provide a more detailed examination of certain ML biases and mitigation methods (e.g., Baeza-Yates, 2018; Mehrabi et al., 2019; Silva & Kenney, 2019; Suresh & Guttag, 2019), and across these articles the respective terminologies differ substantially. We unite these dispersed perspectives with a systematic, interdisciplinary review that summarizes potential ML biases throughout the full ML project lifecycle.

3. Conceptual framework and research process

3.1. Process phases of machine learning projects

Fig. 2 displays the six phases of the CRISP-DM process model that can be used to plan, organize, and implement an ML project (Martínez-Plumed et al., 2019). Anticipating our findings, we also embed the eight ML biases from our review into this figure. Published in 1999 to standardize data mining processes across industries, the CRISP-DM has since become the most common process model for data mining, data analytics, and data science projects.

The initial *business understanding* phase focuses on understanding the ML project objectives and requirements from a business perspective, then converting this knowledge into an ML problem definition and a preliminary plan designed to achieve the objectives. The *data understanding* phase starts with initial data collection and proceeds with activities that enable the researcher to become familiar with the data (i.e., describe data, explore data, and verify data quality). The *data preparation* phase covers all activities needed to construct the final dataset from the initial raw data that will be used for the ML algorithm. Tasks include observation and attribute selection as well as transformation and cleaning of the data. Several ML techniques are then selected in the *modeling phase* and applied to the prepared dataset, before their parameters are calibrated to optimal values. The *evaluation phase* determines which ML model best meets the success criteria and reviews the work accomplished. Before proceeding to the final deployment of the ML model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain that the ML model adequately achieves its objective. Depending on the requirements, the *deployment phase* can be as simple as generating a report or as complex as implementing a self-learning algorithm for organizational decisions (e.g., marketing spending per channel).

3.2. Data collection and analysis

We conducted a systematic, problem-centered literature review to integrate existing knowledge about ML biases through the conceptual lens of the CRISP-DM model. First, different types of ML biases are identified and consolidated into distinct categories. Second, possible mitigation methods that address these biases are grouped. Third, both ML biases and mitigation methods are incorporated into the different phases of the CRISP-DM model.

To identify relevant articles for our review, we performed a systematic keyword search in EBSCO, AIS Electronic Library, ACM Digital Library, ScienceDirect and Emerald Data Base with a focus on leading journals in Business, Marketing, and Information Systems research. This initial search revealed 61 articles in total. We then conducted a screening, and only considered articles that are peer-reviewed and/or

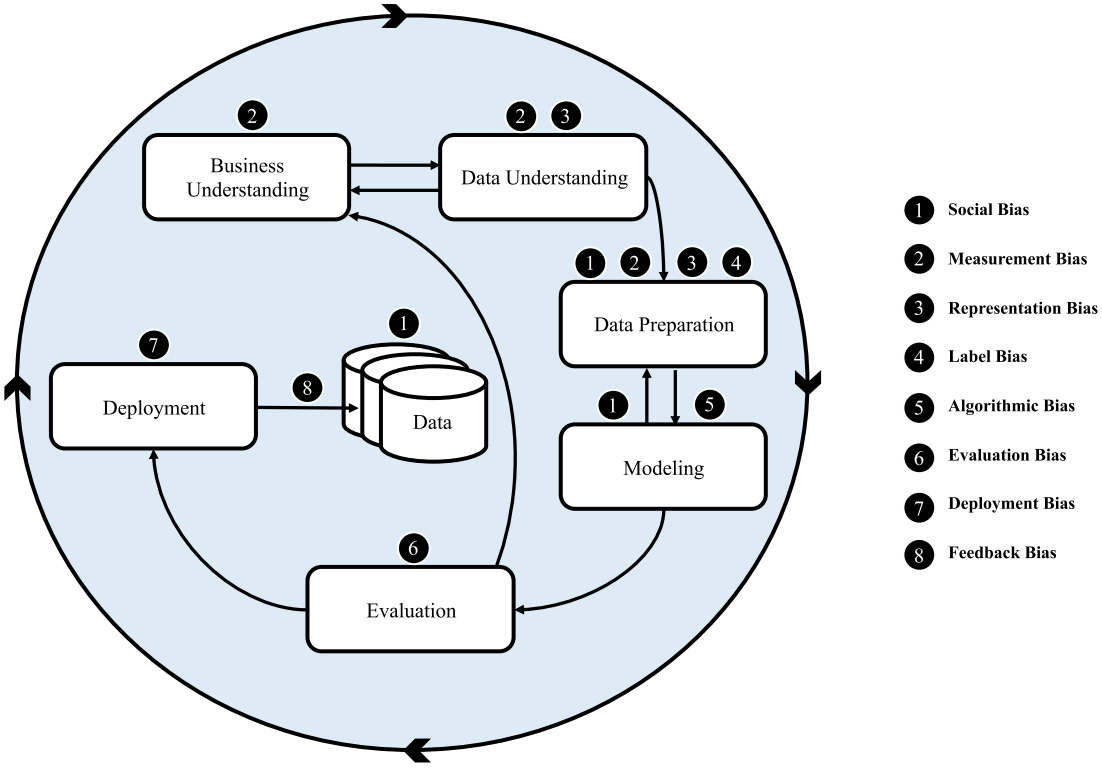

Fig. 2. Process Phases and Potential Biases of Machine Learning Projects Note: The cross-industry standard process for data mining process model is based on Wirth and Hipp (2000).

conference proceedings, define at least one relevant ML bias, and explain at least one relevant identification or mitigation method. After applying these criteria, 24 relevant articles remained. We next performed forward and backward search on the relevant articles using Web of Science. This step added 31 articles based on the same inclusion/exclusion criteria. Finally, senior scholars in the research domain were contacted for additional relevant literature, and 13 articles were added. In total, 68 articles were included in the literature review. We provide an overview of all articles in the appendix. The literature search process is depicted as a flowchart diagram in upper part of Fig. 3.

In the data analysis phase, the different biases and mitigation methods were extracted from the articles. Because different synonyms

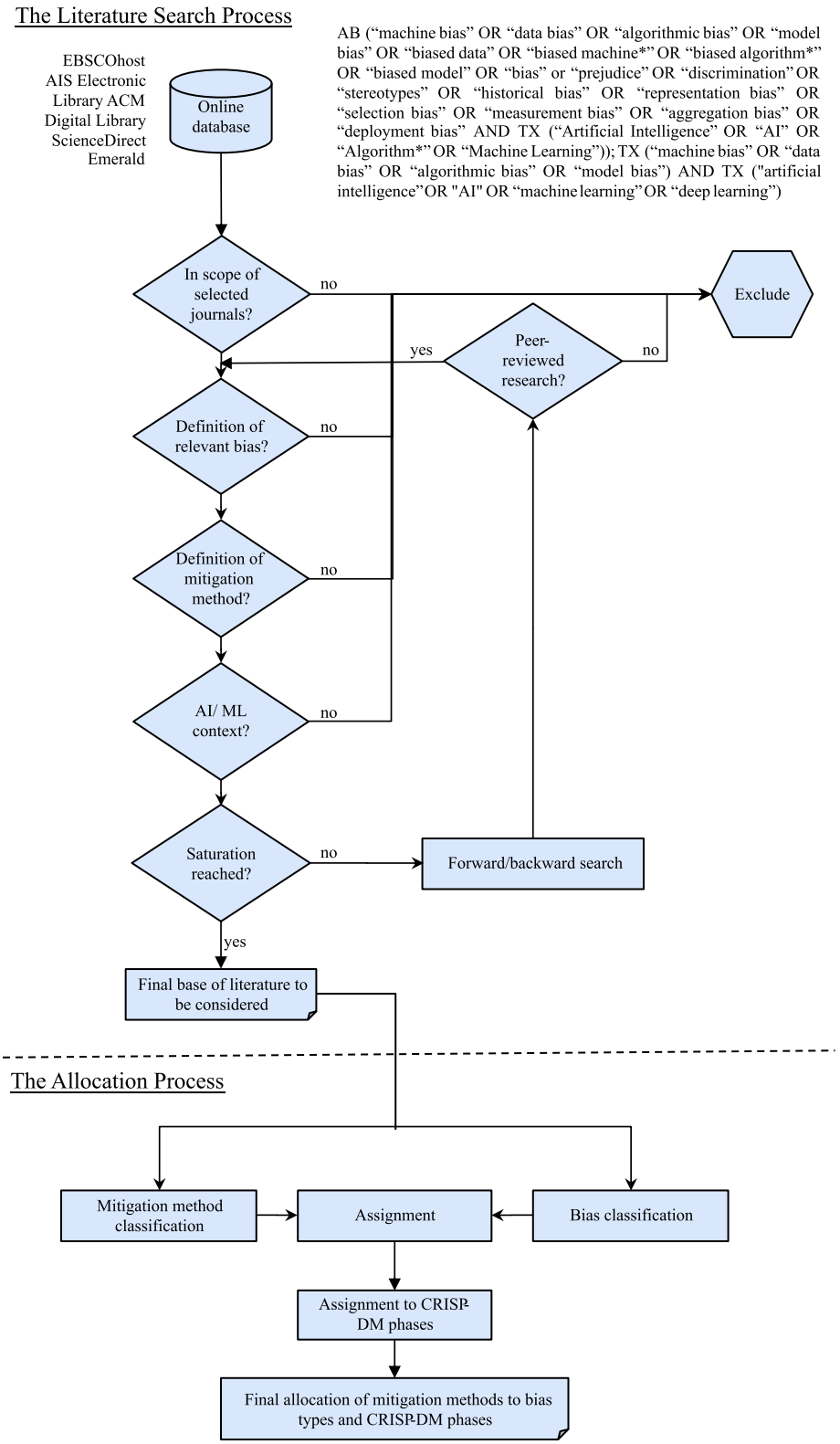

Fig. 3. Selection Criteria and Evaluation Framework.

exist in the literature for the same type of bias, all biases were then descriptively synthesized based on their mechanism. This allowed us to code the biases into eight distinct categories, as summarized in Fig. 2: social bias, measurement bias, representation bias, label bias, algorithmic bias, evaluation bias, deployment bias, and feedback bias. To derive the eight distinct biases, we followed the method for taxonomy development by Nickerson et al. (2013). Specifically, we used the empirical-to-conceptual approach meaning that we started with empirical data clusters and deductively conceptualized the nature of each cluster afterwards. The *meta*-characteristic used for distinguishing the identified biases is the origin (i.e. the occurrence in the CRISP-DM process model) and cause of the bias (i.e. in the data, through humans, in the algorithm, etc.). We stopped iterating when the objective (no bias was merged with a similar bias or split into multiple biases in the last iteration) ending conditions were met.

The taxonomy meets subjective ending conditions as well. The taxonomy is concise, robust, comprehensive, extendible, and explanatory. With eight dimensions, it is meaningful without being overwhelming (concise). The dimensions can differentiate among objects (robust) and classifying random samples of objects within the domain (comprehensive). New dimensions can easily be added if needed (extendible). As the identified characteristics are sufficient to understand the differences among types of bias, the taxonomy is also explanatory. The eight distinct biases were then assigned to the phases of the CRISP-DM process model by matching the mechanism that leads to a specific bias to the tasks in each of the phases (see Fig. 2).

We also identified 24 mitigation methods, allocated them to the respective bias they address, and assigned them to the phases of the CRISP-DM process model. The allocation of mitigation methods was independently conducted by two researchers to enhance reliability. This process involved the extraction and description of each identified mitigation method and an assessment of how the method was suited to addressing the identified types of bias. In most of the analyzed articles the bias-mitigation method relationship was evident. For each mitigation method, and particularly in case of ambiguities, the research team carefully reviewed how each mitigation method affected elements in the CRISP-DM process and how it mitigated bias. In the next step, each bias mitigation method was assigned to a particular process phase in CRISP-DM. The lower part of Fig. 3 provides an overview of the allocation process.

4. Machine learning biases and mitigation methods

4.1. Overview of machine learning biases

We start by describing the eight distinct ML biases summarized in Table 1. A *social bias* occurs when available data reflects existing bias in the relevant population prior to the creation of the ML model. When data embodies a social bias, the resulting ML model will most likely lead to unwanted outcomes. Even if the data is perfectly measured and sampled, a normative concern with the current state of the relevant population may exist that should not be reinforced by the ML model (Mehrabi et al., 2019; Obermeyer et al., 2019; Olteanu et al., 2019). For example, statistical evaluations revealed in 2018 that only 5% of the Fortune 500 companies' CEOs were women. This unequal distribution was consequently reflected in Google's image search of CEOs that showed only a small fraction of women. Google has recently adapted the search results on images of CEOs showing a higher proportion of women in order to not reinforce gender inequality (Suresh & Guttag, 2019).

A *measurement bias* is introduced if chosen features and labels are imperfect proxies for the real variables of interest. When defining the target variable and necessary features for the ML problem, researchers may choose imperfect proxies for the true underlying value or include protected attributes (Mullainathan & Obermeyer, 2017). Protected attributes refer to attributes such as race, gender, or ethnicity that partition a population into different groups that should be treated equally.

Table 1
Overview of Machine Learning Biases.

Bias	Definition	Synonyms	Selected References
SocialBias	Available data reflects existing bias in the relevant population prior to the creation of the ML model.	Historical Bias, Societal Bias, Individual Bias, Pre-existing Bias	Mehrabi et al., 2019,Obermeyer et al., 2019, Olteanu et al., 2019
Measurement Bias	Chosen features and labels are imperfect proxies for the real variables of interest.	Linking Bias, Omitted Variable Bias	Mullainathan & Obermeyer, 2017Suresh & Guttag, 2019
Representation Bias	The input data is not representative for the relevant population, which leads to systematic errors in ML model predictions.	Temporal Bias, Longitudinal Data Fallacy, Emergent Bias, Population Bias, Group Bias, Aggregation Bias, Behavioral Bias, Sampling Bias, Content Production Bias, (Self) Selection Bias, Availability Bias	Baer, 2019, Barocas & Selbst, 2016, Lan et al., 2010, Olteanu et al., 2019
Label Bias	Labelled data systematically deviate from the underlying truth categories.		Barocas & Selbst, 2016, Olteanu et al., 2019
Algorithmic Bias	Inappropriate technical considerations during modeling lead to systemic deviation of the outcome.	Statistical Bias, Technical Bias	d'Alessandro et al., 2017, Friedman & Nissenbaum, 1996, Mehrabi et al., 2019
Evaluation Bias	A non-representative testing population or inappropriate performance metrics are used to evaluate the ML model.	Observer Bias, Funding Bias	Mehrabi et al., 2019, Olteanu et al., 2019
Deployment Bias	The ML model is used and interpreted in a different context than it was built for.	Cause-Effect Bias	Mehrabi et al., 2019, Olteanu et al., 2019
Feedback Bias	The outcome of the ML model influences the training data such that a small bias can be reinforced by a feedback loop.	Presentation Bias, User Interaction Bias, Popularity Bias, Ranking Bias, Second Order Bias	Mehrabi et al., 2019, Olteanu et al., 2019

Using protected attributes as proxies for other features of interest may result in a discriminant or inaccurate classifier. But even if the protected attribute is excluded, the discriminant effect can still exist due to the redlining effect, which states that protected attributes can correlate with non-protected attributes and still bias the outcome (Corbett-Davies & Goel, 2018; d'Alessandro et al., 2017). For example, in a crime prediction application, the feature "number of arrests" is used to predict future criminal activity. Assuming African American and Caucasian defendants commit the same number of drug sales, they have a similar true risk. However, in minority neighborhoods with heavier policing, African American defendants are more likely to experience drug arrests. Despite the similar true risk, the ML model would therefore classify African Americans as a higher risk than Caucasians (Angwin et al., 2016).

A *representation bias* arises when the input data is not representative

of the relevant population, which leads to systematic errors in model predictions. This bias relates to a lack of representability either evoked by the specific time of data collection or the specific sample. First, a representation bias can occur when the training data is no longer representative of the relevant population when the model is deployed. For example, data can be disturbed by one-time phenomena. ML algorithms built for credit card applications use historical data regarding the probability of a credit default. In case of an unsuspected event during the data collection, such as a natural catastrophe in a certain area, people might not be able to pay back their debts. Therefore, applicants from this area will most likely be classified as potential defaults (Baer, 2019). Second, a representation bias can also occur when the training data is wrongly sampled. Such a bias emerges if the distribution of the sampled population differs from the "true" underlying distribution in the relevant population. This over- or underrepresentation can have several causes, including difficult or expensive availability of required data. The algorithm consequently fails to make good predictions for the overall population (Lan et al., 2010).

A *label bias* arises when training data is assigned to classes or labels that systematically deviate from the underlying truth categories. Researchers often face the difficulty of deciding which available label best applies to the present data, and existing labels may also fail to precisely capture meaningful differences between classes (Barocas & Selbst, 2016). Moreover, due to ambiguity and cultural or individual differences, labels might systematically deviate. One such example is the assumption that a certain number of pictures are to be labeled as "wedding". A person that is brought up in western culture will likely only label pictures with brides in white dresses and grooms in dark suits as "wedding", and thus fail to label pictures of an Indian wedding, with its colorful dresses and special decorations, as a wedding (Baer, 2019).

An *algorithmic bias* occurs when inappropriate technical considerations during modeling lead to systemic deviation of the outcome. This occurs when formulating the optimization problem when researchers make data and parameters amenable to computers (Dwork et al., 2011; Friedler et al., 2019). The resulting ML model may fail to treat groups fairly when the probability of misclassification, i.e., false-positive and false-negative rates, are distributed unequally among groups (Bellamy et al., 2018; d'Alessandro et al., 2017). For example, minorities exhibited a higher false-positive rate than majority groups in COMPAS, a predictive policing application that assesses the risk of crime recidivism (Chouldechova, 2017). Such a misclassification can also result in service discrimination for certain individuals (Ukanwa & Rust, 2020).

An *evaluation bias* takes place when a non-representative population or inappropriate performance metrics are used to evaluate the model. ML models are often tested on the same benchmark data to allow for an objective comparison. However, if the benchmark itself is not representative, models could be preferred that only perform well on a subset of the relevant population (Suresh et al., 2018). Thus, choosing the wrong benchmark data can lead to overlooking potential biases. For example, if a facial recognition algorithm is trained on a dataset with underrepresented dark-skinned females and is tested on a similarly unbalanced benchmark, the bias will remain unrecognized (Suresh & Guttag, 2019).

A *deployment bias* arises when the model is used and interpreted in a different context than the one it was built for. This bias occurs because no ML algorithm operates fully autonomously; rather, an ML algorithm requires the input of human decisions when it is deployed. The researchers deploying an algorithm may differ from those that built it: They may have a different knowledge base or values and interpret the algorithmic output according to their internalized biases (Bellamy et al., 2018; Chouldechova, 2017). For example, certain risk assessment ML models are built to predict the likelihood of a criminal committing a future crime. However, in practice, these models are often used in different contexts, such as determining the length of defendants' sentences (Collins, 2018).

A *feedback bias* arises when the outcome of the ML model influences the training data such that a small bias can be reinforced by a feedback loop. It emerges when the output of the ML model is used as a new input, and the algorithm is refined over time (e.g., through re-training). If the outcome of the ML model has an influence on subsequent training data, an initially small bias may be potentially reinforced through a feedback loop (Bellamy et al., 2018; Martin, 2019). For example, once a certain piece of content attains a good ranking according to a rating algorithm based on the number of times it has been clicked, it will affect the position and the promotion of this content, thus leading to even more clicks. Consequently, a reinforcing feedback loop is created and can lead to decreased user satisfaction when unwanted content is promoted (Baeza-Yates, 2018).

4.2. Overview of mitigation methods

We also identified 24 mitigation methods for addressing the aforementioned biases within the CRISP-DM process phases, as summarized in Table 2. Notably, a particular bias can be mitigated by several methods, and a particular method can mitigate multiple biases. In addition, a mitigation method that is applied in one phase can address biases that occur in the respective phase or in the later stages of the ML project.

4.2.1. Business understanding phase

Three methods prevent the emergence of biases by understanding the business objectives and undertaking actions to ensure a precise translation into ML problems. First, *setting up a diverse research team* helps to mitigate measurement bias, and prevents representation and deployment bias occurring in the data preparation and deployment phases. Diverse teams can identify potential harms by introducing different perspectives on the ML task. This enables teams to better define the ML problem with more appropriate features, specify representative populations, and anticipate different use contexts (Barocas & Boyd, 2017; Jones, 2019).

Second, *exchanging with domain experts on project objectives* addresses emerging measurement bias and prevents representation bias in the data preparation phase. The interaction with domain experts helps to design the ML model with appropriate and measurable target variables and features as well as to consider all possible affected populations (Baer, 2019; d'Alessandro et al., 2017).

Third, *discussing technical and social consequences* of the ML model prevents deployment bias. A researcher should envision the respective social context and consider prevailing moral values (Friedman & Nissenbaum, 1996; Martin, 2019). In addition, constraints regarding the applications on other use contexts should be clearly articulated (Buolamwini & Gebru, 2018).

4.2.2. Data understanding phase

Three methods identify and prevent possible biases through a good prior understanding of the data and its underlying relationships in the relevant population. First, it is often necessary to choose proxies for variables of interest in case they are not directly observable. A statistical estimation of *appropriate proxy variables* mitigates the occurrence of measurement bias. Examining the underlying correlations of the proxies and the true variables of interest supports an appropriate proxy selection (Corbett-Davies & Goel, 2018; d'Alessandro et al., 2017).

Second, *data plotting* can reveal spikes (i.e., one-time phenomena) that affect as outliers any empirical conclusions and need to be removed to prevent representation bias (Baer, 2019).

Third, *exchanging with domain experts on data selection* ensures a thorough understanding of the data and proxy variables in question. Domain experts better determine the application context and can recommend features that should be included for model training to mitigate measurement and representation bias. In addition, researchers often face data labeling challenges in the data preparation phase. Gaining insights from domain experts reduces ambiguity in these

Table 2
Mitigation Methods to Address Machine Learning Biases.

Phase	Method	Example from the Case	Recommended Action
Business Understanding	Setting up a diverse research team	The senior management of the bakery might optimize its product portfolio to rather old customers. A team including younger members will point out that younger customers prefer different products and should be considered as well.	Establish diverse teams to introduce different perspectives on the ML task.
	Exchanging with domain experts on project objectives	When giving discounts to customers with the highest expenditure per visit, men could be favored against women: if they frequent the bakery on weekends, they likely have higher expenditures per visit. Such a social bias could be revealed by domain experts.	Exchange with domain experts to include appropriate and measurable features in the ML task.
	Discuss technical and social consequences	A forecast model predicts the average spending per visit of customers and reveals that customers who purchase a coffee have a higher average spending per visit. Hence, a campaign is launched: To increase the average spending, some customers get free coffee. A discussion on how the forecast model that predicts the average spending per visit of a customer can be used when deployed and comparing the purpose of the model with its intended use could mitigate the bias.	Discuss consequences of the use of the ML algorithm in the respective real-world context.
Data Understanding	Appropriate proxy variables	Temperature is a potential proxy for the "beauty" of the weather. But in reality, sales is not dependent on the temperature but on sunshine. By looking at the correlation of temperature and sunshine, it can be found out whether temperature is an appropriate proxy variable.	Examine the underlying correlations of the proxies and the true variables helps selecting variables.
	Data plotting	Plotting sales data from 2020 can reveal a rapid drop in sales in the beginning of march due to the pandemic. Removing the data affected by the pandemic can prevent representation bias that would have emerged otherwise.	Plot data to reveal possible spikes (i.e., one-time phenomena) that need to be removed from the data.
	Exchanging with domain experts on data selection	For an automated baked goods quality control system that uses computer vision, it is of great importance to label the training images correctly. Exchanging with domain experts can help identify labelling issues, such as "badly" baked products being incorrectly labeled as "good" products.	Exchange with domain experts on the application context to identify features that should be included for model training.
Data Preparation	Data massaging	If an algorithm is used to determine which customers should get a voucher and due to the choice of predictors, young people almost never get vouchers, some of these young people that are close to being assigned to the favorable outcome are relabeled to "gets voucher".	Relabel individuals from an unprivileged group to favorable and individuals from privileged groups to unfavorable outcomes.
	Reweighing	The majority of young people that are assigned to the non-favorable outcome ("no voucher") are down-weighted whereas the minority of young people that are assigned to the favorable outcome ("voucher") are up-weighted. A similar procedure is done for the group old people, only vice-versa.	Balance out datasets by up-weighing subgroups with different weights for each combination of group and label.
Data Preparation	Targeted data augmentation	For an automated quality control system that uses computer vision, it is important to have balanced classes. That is, the dataset should include as many images of "good" products as "bad" products. As most baked products are fine, the dataset will be skewed. The number of images of "bad" baked products can be increased by adding slightly modified pictures (e.g., rotating, mirroring) to the data.	Improve the balance of the dataset by populating parts of an underrepresented group in the dataset.
	Rapid prototyping	If a sales prediction algorithm is developed, an important variable for prediction could be sunniness. By rapidly prototyping the model, it could be revealed that the weather forecast, which is used as an input to the model, is not accurate enough and cannot serve as a predictor.	Create a prototype of the ML algorithm and test it in the field.
	Preprocessing algorithms	If an algorithm is used to determine which customers should get a voucher, gender could be a protected variable. Optimized preprocessing transforms the data by trading off discrimination control (the decision on providing vouchers should be as independent of gender as possible), data utility, and individual distortion (a model trained on the transformed dataset is as close as possible to a model that is trained on the original dataset).	Preprocess the data by using several available algorithms, including disparate impact remover, learning fair representation, and optimized preprocessing.
Modeling	Prejudice remover	If an algorithm is used to determine which customers should get a voucher and "young people" almost never get vouchers, prejudice remover adds a regularizer to the loss function of the model. The regularizer increases the loss function during training if the chosen set of model parameters lead to not giving "young people" a voucher. This way, the algorithm is incited to giving vouchers to "young people".	Introduce regularization terms or constraints that consider differences in how the learning algorithm classifies protected and non-protected groups.
	Adversarial debiasing	Due to the choice of variables young people (protected attribute "age") almost never get vouchers. With adversarial debiasing, the algorithm learns whom to grant vouchers. At the same time, the algorithm makes its decisions in a way that the adversary is unable to predict the age of someone who has (not) been granted a voucher by the algorithm.	Maximize accuracy while simultaneously removing identification of protected attributes.
	Multiple models	One model is trained for the protected group "young people" and a different model for the non-protected groups ("all others"). This	Learn two separate models: One for the protected group and one for the non-protected group.

(*continued on next page*)

Table 2 (*continued*)

Phase	Method	Example from the Case	Recommended Action
		way, "age" can no longer influence the model, as all instances in both models are in the same age group. The two models are then combined, and probabilities adjusted, so that the amount of granted vouchers is unaffected compared to a single model approach.	
	Latent variable model	A latent variable model first discovers discrimination-free class labels that are independent of protected attributes (e.g., age for voucher). That is, each training example is assigned a new label. Subsequently, a new model is learned that maximized the probability of the discrimination-free data.	Discover the actual class labels that a dataset should contain if it was discrimination-free.
	Interpretable models	If a sales forecast algorithm is interpretable, it is transparent why the algorithm predicted a certain sales amount and how much the different variables contributed to the forecast.	Foster transparency and trust in algorithmic models to aid identification of biases.
Modeling	Splitting and resampling	If a random training and test set split leads to a training set that contains only Sundays on which business was poor, the algorithm learns a misleading pattern in the data. By randomly splitting the dataset multiple times and iteratively setting the parameters each time, the parameters are not dependent on a single training set.	Use multiple subsets of the training data and test data.
	Equalized odds	Equalized odds ensures that the rates at which an age group gets vouchers when it should get vouchers (true positive rate) and the rate at which the group gets vouchers when it should not get vouchers (false positive rate) is equal across groups. Thus, no age group gets unjustified vouchers more often than other groups.	Ensure that true positive and false positive rates are equal across protected groups.
	Multitask learning	If the predictive quality is different among different products in a demand forecasting setting multitask learning ensures similar predictive accuracy among groups. First, different product groups are identified. Second, an outcome for separate product subgroups is predicted in a multi-task framework, where each subgroup is a separate task.	Parametrize different groups differently and learn simpler, multiple functions to account for group differences.
Evaluation	Representative-ness of the benchmark dataset	A demand forecast model for train station locations is trained on respective sales and then benchmarked with a free reference dataset from the internet. The forecasting model will perform poorly if the benchmark dataset is not representative for the particular location.	Verify that a benchmark dataset contains a balanced composition of all subgroups present in the model.
	Subgroup validity	For a forecasting model that predicts demands of different product groups together performance gaps between subgroups can exist: The predictive quality could be better for some subgroups than for others. Subgroup validity examines performance metrics in greater detail across subgroups.	Compare performance metrics across groups instead of accepting an aggregated metric to reveal performance gaps.
Deployment	Monitoring plan	A monitoring plan for a sales forecast algorithm helps detect drifts in the data. For example, if sales for a certain product slowly decrease over time and the algorithm is not regularly re-trained the sales forecasts gradually get worse over time. Monitoring reveals such errors and ensures the algorithm is still suitable for the changing context.	Account for changes in the algorithm when the context evolves.
	Human supervision	If due to an error in a sales forecast algorithm an influential input variable like "sales of previous day" is unusually high, the forecast will also be far too high. This would lead to economic harm, if said forecast would directly be transformed into a production order.	Include humans in the application loop to analyze and question algorithmic recommendations.
	Randomness	If a sales forecast algorithm predicts too few items of a product for the next day, this product will also be sold fewer because it will be sold out very early. This will further influence the prediction of the following day and decrease the forecasted amount. Thus, a downward spiral is entered. Randomly increasing the forecasted amount of a product gives customers the chance to buy more items so that the algorithm increases its predictions again.	Lower the impact of the ML model on data generation or sampling distribution.

decisions and prevents label bias (Barocas & Selbst, 2016; d'Alessandro et al., 2017).

4.2.3. Data preparation phase

Five methods mitigate biases by modifying the data prior to the modeling. First, *data massaging* mitigates social bias by strategically relabeling data points near the classification margin according to a ranking of the class probabilities. For instance, by relabeling individuals from an unprivileged group to favorable outcomes and simultaneously individuals from privileged groups to unfavorable outcomes, discrimination can be reduced while maintaining the overall class distribution (Kamiran & Calders, 2012).

Second, with *reweighing* it is possible to address social bias and representation bias. Unrepresentative datasets are balanced out by upweighing underrepresented subgroups with different weights for each combination of group and label. With this approach, discrimination can be significantly reduced while overall class probability is maintained (Hajian & Domingo-Ferrer, 2013; Kamiran et al., 2013).

Third, *targeted data augmentation* reduces representation bias by populating parts of the underrepresented group in the dataset (Chen et al., 2018).

Fourth, *rapid prototyping* is an effective approach for identifying different types of unintended bias (Friedman & Nissenbaum, 1996). By creating and testing a prototype of the ML model, researchers can reveal discriminative effects resulting from social bias, test variables and proxies regarding their suitability to predict the outcome of interest to address measurement bias and uncover overlooked sections of the population to prevent representation bias.

Fifth, *preprocessing algorithms* transforms data and mitigates social bias. For example, a disparate impact remover edits features and labels

in the data by learning a probabilistic transformation and applying rank ordering within groups. This ensures that information related to the non-protected attributes is preserved and the class can still be correctly predicted (Friedler et al., 2019). Learning fair representation formulates an optimization problem of finding an intermediate representation of the data that encodes it well but simultaneously removes information about membership of a protected group. The new representation space captures true underlying features that differ across groups and can then be used to learn a new classifier in the modeling phase that does not use group information (Zemel et al., 2013). An optimized preprocessing formulates a (quasi-)convex problem for the transformation and edits features and labels while complying with fairness constraints (Calmon et al., 2017).

4.2.4. Modeling phase

Six model-based methods conduct modifications of learning algorithms to mitigate biases, and two additional, non-model based methods are available to address biases. First, a *prejudice remover* is an approach that introduces regularization terms or constraints that mitigate social bias. It considers differences in how the learning algorithm classifies protected and non-protected attributes such as race, gender, or ethnicity, and then penalizes the total loss based on the amount of the difference (Kamishima et al., 2012; Zafar et al., 2015).

Second, *adversarial debiasing* maximizes accuracy while simultaneously reducing the ability to identify protected attribute(s). The outcome does not carry any group discrimination information, which helps to mitigate social bias during classifier training (Zhang et al., 2018).

Third, *multiple models* can be used for Naive Bayes Classifiers to learn two separate models, one for the protected group and one for the non-protected group (e.g., contrasting males and females). This way, the protected attribute, as well its proxies, no longer influence the outcomes of the separate models. After combining both models, probabilities are modified so that the distribution of labels is kept similar to the original dataset (Calders & Verwer, 2010).

Fourth, a *latent variable model* can discover the actual class labels that a dataset should contain if it were discrimination-free. The parameters of the model are then set in such a way that the likelihood of the dataset is maximized (Kamiran & Calders, 2009).

Fifth, the design of *interpretable models* fosters transparency and trust in algorithmic models and aids identification of algorithmic biases (Binder et al., 2016; Corbett-Davies & Goel, 2018; Martin, 2019).

Sixth, *splitting and resampling* the training and test data helps to build a robust classifier and consequently mitigates algorithmic bias. It also prevents bias in the evaluation phase by improving diversity in the test set (Berardi et al., 2004; Friedler et al., 2019; Lan et al., 2010).

In addition, two post-processing methods can be applied after the algorithmic training. First, *equalized odds* mitigates social bias by accessing only aggregated data. It solves a linear problem that finds probabilities with which to change and equalize differences in output labels (Corbett-Davies & Goel, 2018; Hardt et al., 2016).

Second, *multitask learning* is an efficient decoupling technique that learns different classifiers for different groups, thereby mitigating algorithmic bias. It parametrizes different groups differently and learns simpler, multiple functions to account for group differences (Dwork et al., 2017; Suresh et al., 2018).

4.2.5. Evaluation phase

A possible evaluation bias can be addressed using two methods. First, the *representativeness of the benchmark dataset* should be verified regarding its balanced composition of all subgroups present in the relevant population (Ryu et al., 2017).

Second, the *subgroup validity* approach compares performance metrics across groups instead of accepting an aggregated metric. If performance gaps between different subgroups are revealed, data augmentation can balance data of underrepresented subgroups to improve overall validity (Buolamwini & Gebru, 2018; Mitchell et al., 2018; Suresh & Guttag, 2019).

4.2.6. Deployment phase

Three methods prevent deployment bias and feedback bias. First, a *monitoring plan* accounts for changes in the algorithm when the context evolves (Corbett-Davies & Goel, 2018; Friedman & Nissenbaum, 1996).

Second, algorithmic recommendations should not be accepted "blindly" because they cannot be expected to be bias-free. Including *human supervision* in the deployment to analyze and question outcomes enhances objectivity, mitigates possible occurrence of deployment bias, and prevents feedback bias (d'Alessandro et al., 2017).

Third, if the outcome of an ML model has an impact on subsequent data generation or the sampling distribution, *randomness* can be introduced to lower the impact of an ML model and thus prevent a feedback bias (Baer, 2019).

5. Machine learning biases in a marketing context

We next use a single, in-depth case study to illustrate the previously identified ML biases and mitigation methods in a marketing context. The illustration in a plausible and naturalistic setting should guide researchers and practitioners in avoiding and mitigating these biases. The company of interest is a nationwide bakery chain, which operates centralized production facilities and multiple bakeries in two distinct location types, either in city centers or at train stations. The company uses ML models for decision-making regarding demand forecasting, promotions and campaigning, new product development, and their loyalty program. One of the authors advises the company on their ML projects and has frequently interacted with the management and data scientists, and was granted full access to their data, algorithms, and decision outcomes. Nevertheless, to ensure the anonymity of the company, we describe fictional biases based on real-world examples and experiences. We use the conceptual model depicted in Fig. 1 to denote the relevant population (i.e., all potential customers of the bakery), and use the data that is generated from the population to train the ML model, which generates predictions that trigger marketing decisions and actions, in turn affecting the relevant population. We next illustrate the eight ML biases by outlining their marketing context and appropriate mitigation methods (see Fig. 4).

5.1. Social bias: A customer reward initiative that ignores loyal customers

For the bakery's 10th anniversary, the marketing team initiated a reward initiative in which particularly loyal customers were to receive a voucher. In determining the criteria for the reward scheme, a classification algorithm was applied to the bakery's customer sales data. It transpired that the features "average spending" and "frequency of visits" were important for the classification of customers who generate most revenue and should therefore be rewarded. Due to the classification, mainly business travelers were rewarded because of their relatively high spending and frequency of visits. However, they are not very loyal in reality. In contrast, teenagers from nearby schools who made frequent visits but had lower spending were disadvantaged even though they are very loyal customers.

The bias emerged because the data-driven process neglected the fact that teenagers were disadvantaged in the reward initiative compared to higher-earning business travelers because of their relative lack of purchasing power. Although the data used for classification led to an economically and mathematically correct classification of customers, it is questionable whether it was socially desirable to disadvantage teenagers who are loyal customers. Such a bias also entails reputational risk and may possibly jeopardize future business potential since teenagers represent a customer group with increasing customer value as they grow older. An effective method to counter such a social bias is rapid prototyping. The bakery should have deployed the ML-model for a short

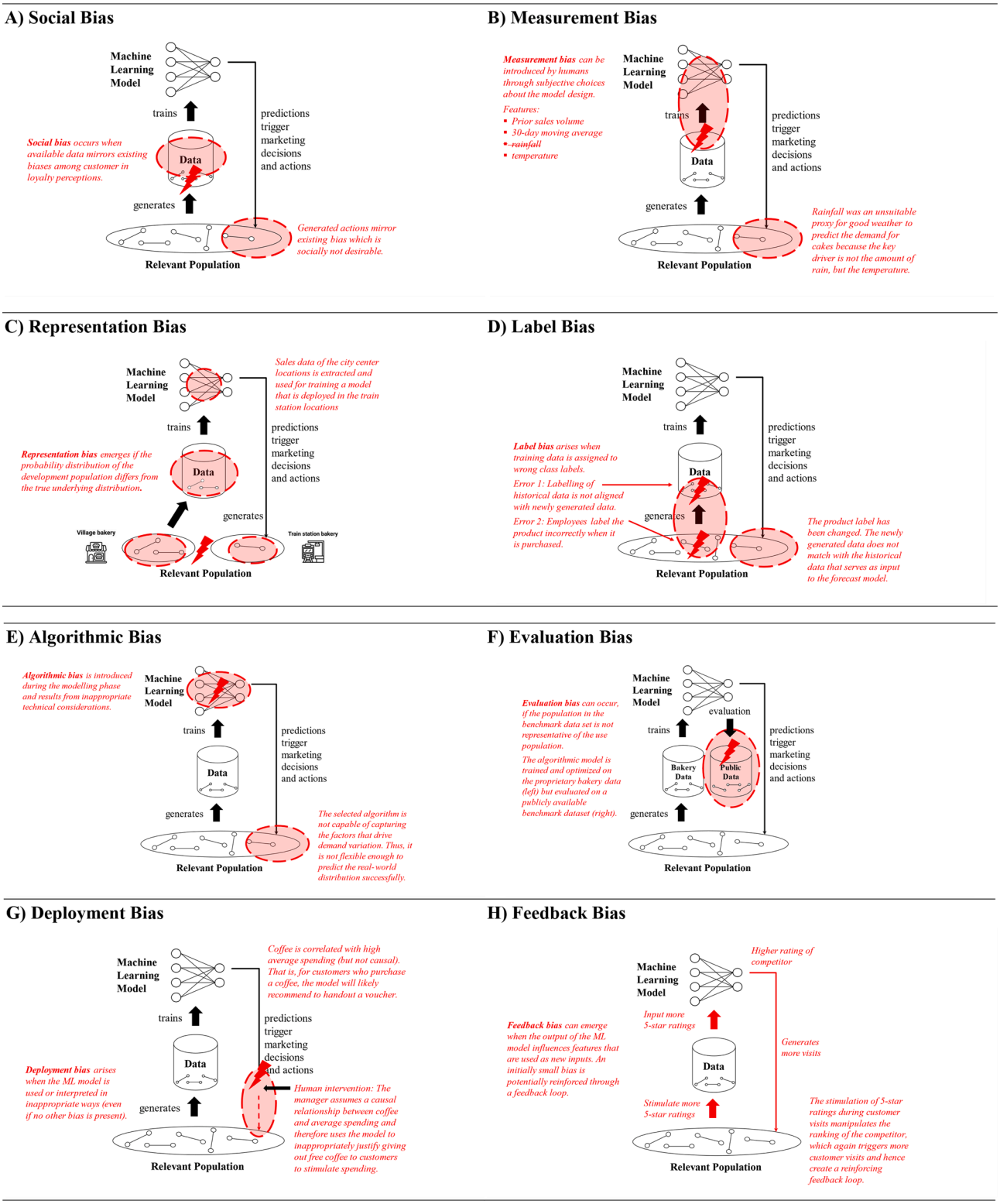

Fig. 4. Illustrations of Machine Learning Biases.

period of time only and asked the bakery staff for feedback on the selection of customers. Another way of mitigation would have been if the staff had handed out vouchers to their most loyal customers to compare the ML-generated selection with the experience-based selection of customers.

5.2. Measurement bias: Capturing good weather effects

Cakes have a high margin for the bakery but also a limited shelf life. Consequently, production and sales quantities must be carefully planned, as both over- and under-capacities are costly (i.e., they result in loss of sales or discarded products). The management knew from experience that more cakes were sold in summer than in winter and suspected that

this was related to the fact that it is sunnier in the summer months. When building the demand forecasting model for cakes, the bakery selected a number of features such as the sales volumes in previous years, the 30-day moving average, and the amount of rainfall forecasted for the next week.

However, it turned out that rainfall was an unsuitable proxy for "good weather" and hence unsuitable to predict the demand for cakes. In fact, customers bought cakes even in rainy weather. The key driver, in fact, was not the amount of rain but the temperature. More cakes were bought on warm days (even when it was raining) than on cold days (even when it was not raining). One effective method to counter such a measurement bias is exchanging with domain experts who often have a more intuitive feeling about the real cause. A production manager who estimates the expected sales amount every day could have helped to find the "real" proxy for good weather and could have prevented the bakery from choosing a wrong proxy.

5.3. Representation bias: Using apples to predict oranges

The bakery trained a demand forecast model with sales data from city center locations. The algorithm learned the demand patterns and temporal distribution for these specific locations. The same model was then used to create a forecast for train station locations which performed very poorly. The dataset for building the forecast model is not representing the demand distribution as customers in the city center locations and travelers have different consumption patterns. For example, demand in city center locations is highest on weekends.

One effective measure to counter representation bias is using interpretable models for which the relationship between input features and output is transparent and explainable. This would have disclosed the strong positive correlation between the input variable "weekend" and the expected demand for the model that was trained on the city center sales data. Together with consulting domain experts, who know that the train station locations sell more on business days, this mitigation strategy would have helped to demonstrate that the probability distributions differ.

5.4. Label bias: Old wine in new bottles

The bakery relabeled an existing product to address the trend towards healthy nutrition. To do this, it re-invented the existing "organic wheat bread" as "organic wellness bread" and created a new product ID. The promotion campaign was completed and the new product was launched on a Wednesday morning, and the organic wellness bread was positioned as fresh and healthy amongst bakery customers. Label bias would be introduced if employees working on Monday, Tuesday, and Friday were aware of the ID change of the newly launched organic wellness bread, but employees working on Wednesday and Thursday were not.

That is, when purchases were entered into the cash register on Wednesday and Thursday, "organic wellness bread" could be incorrectly labelled as "organic wheat bread" in the data. The demand forecast for "organic wellness bread" would be calculated incorrectly. An effective measure to counter label bias is data plotting. If the data is plotted, changes in the ID of the same product can be detected. On a specific day, the sales of the old ID drops to zero, whereas sales of the new ID has approximately the value that the old ID had before. Moreover, possible spikes in the data that result from human errors can be revealed by data plotting.

5.5. Algorithmic bias: As simple as possible, but no simpler

A newly hired sales manager of the bakery selected a forecasting model that proved successful in her former role in a different industry. She was convinced that a 3-day moving average was simple to use, transparent to understand, and easy to communicate, and hence would serve as a suitable algorithm for forecasting the demand for organic wheat bread. However, given lower demands at the beginning of the week and higher demands towards the weekend, the forecasting algorithm did not account for the complexity of the prediction problem.

This algorithmic bias was costly, as the bakery produced a surplus of units from Mondays to Wednesdays and lost sales from Thursday to Saturday. One effective measure to counter algorithmic bias is rapid prototyping. This way, high deviations between the algorithmic performance in the training data and test data can be detected prior to deploying the model. In the example above, the performance will be poor since it is a case of underfitting. Domain experts can suggest additional input features to increase the complexity of the model.

5.6. Evaluation bias: Benchmarking with caution

The data scientist of the bakery was asked to build a demand forecast model using the sales data from the bakery. Once the prototype for the ML model was developed, a general manager asked for an objective benchmark to develop a better sense of the quality of the prediction model. Because the manager has learned that several online platforms offer free reference datasets, the data scientist is asked to benchmark the developed model with a "neutral" dataset. The data scientist benchmarked the forecasting model against a publicly available bakery sales dataset. Unfortunately, the forecasting model performed poorly and the manager refused to provide a budget for the further development of the ML model.

The data scientist could not demonstrate the value of the forecasting model convincingly because it had been specialized on the proprietary dataset but was evaluated against a public dataset that is not representative for the bakery. To mitigate evaluation bias, the representativeness of the benchmark dataset should have been verified with respect to the relevant population of the bakery. The benchmarking result is only meaningful when the representativeness has been established. Only then does an ML model which performs well on the benchmark dataset also perform well in the context where it is deployed.

5.7. Deployment bias: Stick to the knitting

The marketing manager of the bakery launched a campaign that issued vouchers to a specific regional target group. He built a forecast model that predicted the average spending per visit of a customer. When analyzing the model, he learnt that customers who purchased a coffee had a higher average spending per visit. Using this insight, he launched a new campaign: Some customers receive one free coffee to increase their average spending. However, the campaign did not increase the average spending of the target customers.

The marketing manager assumed that average spending per visit could be increased by providing free coffee. However, by doing this, the prediction model was interpreted incorrectly. The original purpose of the model was to predict average customer spending per visit. Deploying the model for a different decision and justifying the arbitrary human intervention of giving out free coffee represented a misuse of the model. From a statistical perspective, the example illustrates the common fallacy of confusing correlation with causation. That is, there is a correlation between the purchasing of a coffee and spending per visit, but that does not mean that average spending per visit can be increased by providing free coffee.

One effective measure to counter deployment bias is to discuss technical and social consequences. In the example above, discussing how the forecast model that predicts the average spending per visit of a customer can be used when deployed and comparing the purpose of the model with its intended use mitigates the bias. Consequently, the marketing manager is prevented from using the forecast model in an inappropriate context. Another applicable measure is establishing a monitoring plan to track whether the intervention of providing free coffee is effective in increasing the average spending. This would reveal

Table 3
General Implications and Actionable Recommendations to Avoid Machine Learning Bias.

General Implication	Actionable Recommendation
There is no bias-free MLAs bias can occur in each phase of the CRISP-DM process model and can be caused by people, data, algorithms, and application context, it should be proactively addressed and mitigated using the full range of mitigation methods.For bias-free ML, a large range of assumptions would need to be true, which is very unlikely. Therefore, there is no guarantee to discover bias in ML models.	• Evaluate and proactively address bias risks of envisioned ML applications. • Complement technical with non-technical mitigation methods to account for social, technical, and economic aspects of ML bias. • Document assumptions and decisions made about the ML application and establish processes to discover bias proactively during development and operation to help mitigate risks from ML bias.
There is no panacea for ML biasesML differs significantly from traditional information systems. One phenomenon unique for ML is the existence of bias which can occur throughout the entire ML project. Bias is mostly introduced unintentionally, and often difficult to detect.The context sensitivity of bias emerges from phenomena that are unique for ML such as high data dependency and adaptivity. This can lead to bias in unexpected situations and requires special measures for mitigating bias risks from ML as compared to traditional information systems.	• Be aware and alert to emerging biases, obtain stakeholder feedback, and be open to suggestions and concerns regarding the ML application. • Establish a process for escalating potential harm resulting from ML models and track changes in social norms. • Develop the capability to use the full armory of bias mitigation methods (technical- and non-technical).
Change triggers ML biasesBias can also occur during operation of the ML model through re-training, feedback loops, or changes of the application context. Consequently, ML applications need to be monitored for biases during their entire operational lifecycle. When an ML model is deployed, employees should be able to understand the underlying procedures of the ML model and hence it is important to make the ML model as transparent and explainable as possible.	• Continuously evaluate the possibility of ML bias, especially driven by changes in the relevant population, its representing data, the ML model, and the deployment context. • Train employees and raise awareness of events and contextual changes that can introduce bias to ML applications and how to monitor and report the model's decisions regarding possible bias. • Co-develop and prototype ML applications with end-users to make the ML model as transparent and explainable as possible.

the ineffectiveness of the model.

5.8. Feedback bias: Mind the power of the algorithm

The demand planning team observed a consistent decline in the number of customers and transactions in the train station location over a period of several months. Seasonal and one-off effects could be ruled out, and the product mix remained similar to prior years. The bakery addressed the decline with a promotion to retain existing and to attract new customers: each customer received a free croissant with every coffee-to-go purchased. Although revenues improved slightly, the overall trend could not be stopped.

Only after speaking to travelers who passed by with a coffee and a croissant from a competitor did the bakery become aware of the problem: The competitor provided a free croissant to each customer who gave the bakery a five-star rating on TripAdvisor and GoogleMaps. Accordingly, travelers would find a competing bakery that had many excellent online reviews and go there. As these travelers would also be rewarded with a croissant for five-star ratings, the competitor would further stimulate the positive online rating and generate a reinforcing feedback loop. The competitor thus understood the power of algorithms and found a way to exploit feedback bias to increase recommendations and customer visits.

As a result, it became increasingly difficult for the bakery to counter the competitive attack as the recommendation algorithm was under the control of third-party platform operators. It is also worth noting that the cost for the reward per campaign interaction (i.e., one croissant) was the same for both bakeries. Nevertheless, mitigation methods help to identify and detect such attacks, e.g., by collecting and monitoring recommendations on relevant platforms in competitive markets, and to initiate countermeasures. These could include mobilizing one's own customer base to generate positive ratings or filing complaints for anticompetitive behavior.

6. General discussion

We conducted a systematic, interdisciplinary literature review and identified eight distinct ML biases (see Table 1), mapped these biases in the CRISP-DM (see Fig. 2), and outlined twenty-four bias mitigation methods (see Table 2). We also proposed a conceptual model to illustrate the typical application logic of ML in marketing (see Fig. 1) and used it jointly with the bias taxonomy to analyze eight ML biases and their effects in a case study of a nationwide bakery chain. This application yielded eight visualizations – one for each identified ML bias – that illustrate the link between each bias cause and effect (see Fig. 4). Based on our work, we distilled actionable recommendations for managers on how to approach ML bias in marketing (see Table 3). Our work contributes to the nascent research on ML in marketing and it holds important implications for researchers and managers alike.

6.1. Contributions to the literature

To date, there is a general scarcity of research on ML bias, despite its high importance. More specific shortcomings include the generic use of the term bias, the inadequately captured link between cause and effect of different biases in the absence of a holistic, comprehensive perspective, and, consequently, challenges in systematically addressing ML bias in marketing through effective mitigation strategies and marketeer education. By addressing these shortcomings we provide several contributions to the literature. First, our comprehensive taxonomy defines the most common biases in ML. Jointly with the conceptual model of the ML application logic, our work provides a focused theoretical account to differentiate eight types of bias. The taxonomy was developed in a stepwise, transparent manner based on a thorough literature analysis. Researchers can use the taxonomy to more effectively study, analyze, and discuss ML bias in other empirical settings, e.g., for further research on organizational practices and processes that generate or address the different types of ML bias, as indicated by Rai (2020).

Second, our results are useful for understanding and analyzing ML bias more systematically, because the developed taxonomy and the respective mitigation strategies provide a holistic view on how ML biases can be identified, avoided, and mitigated along the CRISP-DM development process. CRISP-DM is the most established, and widely accepted standard for managing ML projects. Hence, mapping different bias causes and effects onto CRISP-DM provides a systematic structure for capturing ML bias from a spatial, temporal, and causal perspective. As such, our research addresses Guha et al. (2021, p. 35) call for research by helping to "identify bias […], before much harm is caused", and responds to prior calls for research on the topic (c.f., Davenport et al., 2020).

Third, the examples from of the bakery chain that contextualized the biases in specific marketing ML applications and the visualizations for each bias should be used in marketing education. Since we focused on the logic and circumstances under which ML bias can occur rather than on the mathematical foundations, the illustrative examples can be used to raise marketeers' awareness of potential ML biases. Sensitizing marketing students and practitioners for the challenges associated with the use of ML in real-world marketing applications is an important strategy to avoid ML bias (Huang & Rust, 2021; Puntoni et al., 2021).

6.2. Managerial implications

Besides the concrete recommended actions to address biases in ML projects in each stage of the CRISP-DM process outlined in Table 2, our work also holds some more general implications for managers when considering ML in marketing, as summarized in Table 3.

There is no bias-free ML. All ML applications have (potential) biases, since otherwise the following assumptions would have to be true. First, the available data would need to perfectly represent the relevant population, including every relevant data feature. Second, all characteristics in the relevant population as well as in the data would need to be socially and normatively desired. Third, there would need to be stability in the relevant population and in the data, in order to prevent any changes. Given that these requirements can hardly be met, all ML applications in marketing may suffer from biases. Thus, AI based on ML does not only demonstrate an opportunity but also a risk for automated marketing activities, and researchers and managers should focus on how to proactively address and mitigate the potential of a social bias, measurement bias, representation bias, label bias, algorithmic bias, evaluation bias, deployment bias, and feedback bias.

There is no panacea for ML biases. Our research demonstrates that ML biases constitute a multifaceted problem that has an impact throughout the entire ML project, including the business understanding, data understanding, data preparation, modeling, evaluation, and deployment phases. ML bias is mostly introduced unintentionally and is relatively difficult to detect at face value. Mitigation methods need to address social, technical, and economic aspects of an ML project. As a consequence, technical mitigation methods should be complemented with non-technical mitigation methods that consider more than merely good performance results. In particular, ML bias removal tools such as "AI Fairness 360" (Bellamy et al., 2018) should be assessed with great caution as the promised outcome can hardly be achieved single-handedly with the tool alone. Given the prevalence of human judgment and human errors in ML, researchers and managers should instead use the full armory of mitigation methods provided.

Change triggers ML biases. ML models are largely data-driven and take over some decision-making agency. However, the context in which the decision-making takes place is not stable, but instead constantly evolves and changes, sometimes gradually and incrementally and sometimes dynamically and radically. We suggest that researchers and managers remain reflective regarding the deployment of their ML models and strive to proactively address adverse effects that can result from changes in the relevant population, its representing data, the ML model, and the deployment context. This is especially important in a dynamic discipline like marketing where consumer behavior, technological possibilities, and the legal context frequently change, making ML predictions based on outdated data harmful for companies.

6.3. Limitations and future research directions

Our literature review of ML biases is necessarily "retrospective", restricted to existing articles, and subject to the limitations inherent in the original studies. Thus, we advise that future research addresses some of the limitations of our work. It would be insightful to illustrate the biases and the mitigation methods in specific ML projects with real data. By doing so, the contextual conditions under which each of the identified biases can occur would be better captured. Because there is no one-size-fits-all solution to the diverse ML biases, technical and social aspects of ML should be combined to bring context-awareness to research and practice.

Another future research direction is to identify how human biases potentially translate into ML biases. The starting point for our research was to identify ML biases which are known or have been studied in extant research and how these biases can be addressed using effective mitigation strategies. However, for many, if not most, of the identified ML biases, human bias can be the cause. For example, social bias is introduced into the ML context if data that is used for training a model that resembles such human biases. Also, human bias can lead to representation bias, when humans select a non-representative data set for training. Further biases, which might be stemming from human bias, could be identified and contribute to the nascent stream of research on ML bias. Such kind of studies would likely benefit from decades of research particularly in the psychology domain and the provided CRISP-DM framework, as well as the conceptual model provided here. These prior works could serve as a bridge to translate and locate how and when human bias is salient and how it might take effect within the typical logic of ML applications.

6.4. Conclusion

ML can incorporate inadequate properties that lead to both technically incorrect and socially unacceptable results. Besides performance criteria such as reliability, efficiency, and accuracy, addressing bias should be an integral part of any ML application. In contrast to the dominant discussion on bias, which is primarily about discrimination based on race, gender, religion, or membership in a social minority, we also emphasize and illustrate the economic dimension of bias. As such, we show that managing bias in ML projects is not only about fairness, but also about ensuring sustainable economic value from using ML in business settings.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Ajay Kumar, Dursun Delen, Eric W. T. Ngai, and Stefan Bernritter for their feedback on earlier versions of this manuscript, and Viktoria Huber for assistance in data collection and coding of the machine learning biases and mitigation methods.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). *Machine bias*: Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed on January 6, 2022.

Baer, T. (2019). *Understand, Manage, and Prevent Algorithmic Bias. A Guide for Business Users and Data Scientists*. Apress.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM, 61*(6), 54–61.

Barocas, S., & Boyd, D. (2017). Engaging the ethics of data science in practice. *Communications of the ACM, 60*(11), 23–25.

Barocas, S., & Selbst, A. (2016). Big Data's Disparate Impact. *California Law Review, 104*(3), 671–732.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Berardi, V. L., Patuwo, B. E., & Hu, M. Y. (2004). A principled approach for building and evaluating neural network classification models. *Decision Support Systems, 38*(2), 233–246.

Binder, A., Bach, S., Montavon, G., Müller, K.-R., & Samek, W. (2016). Layer-Wise Relevance Propagation for Deep Neural Network Architectures. *Information Science and Applications (ICISA), 2016*, 913–922.

Bogen, M., & Rieke, A. (2018). *Help wanted: An examination of hiring algorithms, equity, and bias*. Upturn: Technical report.
Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, *81*, 77–91.
Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*(2), 277–292.
Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). In *Optimized Pre-Processing for Discrimination Prevention* (pp. 3992–4001). Curran Associates Inc.
Chen, I. Y., Johansson, F. D., & Sontag, D. (2018). Why Is My Classifier Discriminatory? *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, 3539–3550.
Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data, 5*(2), 153–163.
Cohen, L., Lipton, Z. C., & Mansour, Y. (2019). Efficient candidate screening under multiple tests and implications for fairness. *arXiv preprint arXiv:1905.11361*.
Collins, E. (2018). Punishing Risk. *The. Georgetown Law Journal, 107*, 57.
Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. In *arXiv [stat.ML]*. http://arxiv.org/abs/1808.00023.
Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science, 24*(4), 595–615.
d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data, 5*(2), 120–134.
Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, 108*(17), 6889–6892.
Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48*(1), 24–42.
De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K. U., & von Wangenheim, F. (2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing, 51*, 91–105.
Dwork, C., Immorlica, N., Kalai, A. T., & Leiserson, M. (2017). *Decoupled classifiers for fair and efficient machine learning.* In arXiv [cs.LG]. http://arxiv.org/abs/1707.06613.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226).
Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *In: Proceedings of the conference on fairness, accountability, and transparency* (pp. 329–338).
Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems, 14*(3), 330–347.
Guha, A., Grewal, D., Kopalle, P. K., Haenlein, M., Schneider, M. J., Jung, H., Moustafa, R., Hegde, D. R., & Hawkins, G. (2021). How artificial intelligence will affect the future of retailing. *Journal of Retailing, 97*(1), 28–41.
Hajian, S., & Domingo-Ferrer, J. (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering, 25*(7), 1445–1459.
Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 3315–3323.
Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science, 49*(1), 30–50.
Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems, 28*(1), 3–16.
Kamiran, F., & Calders, T. (2009). Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication* (pp. 1–6).
Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33.
Kamiran, F., Žliobaite, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems, 35*(3), 613–644.
Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. *Machine Learning and Knowledge Discovery in Databases*, 35–50.
Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science, 65*(7), 2966–2981.
Lan, J., Hu, M. Y., Patuwo, E., & Zhang, G. P. (2010). An investigation of neural network classifiers with unequal misclassification costs and group sizes. *Decision Support Systems, 48*(4), 582–591.
Liu, L., Dzyabura, D., & Mizik, N. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science, 39*(4), 669–686..
Luca, M., Kleinberg, J., & Mullainathan, S. (2016). Algorithms need managers, too. *Harvard Business Review, 94*(1), 97–101.
Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing, 37*(3), 481–504.
Mariani, M. M., Perez-Vega, R., & Wirtz, J. (2021). AI in marketing, consumer research and psychology: A systematic literature review and research agenda. *Psychology & Marketing.*, 1–22.
Martin, K. (2019). Designing Ethical Algorithms. *MIS Quarterly. Executive, 18*(2), 129–142.
Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández Orallo, J., Kull, M., Lachiche, N., Ramírez Quintana, M. J., & Flach, P. A. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.
Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2018). Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. In *arXiv [stat.AP]*. http://arxiv.org/abs/1811.07867.
Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review, 107*(5), 476–480.
Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems, 22*(3), 336–359.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.
Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*, 13.
Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing, 85*(1), 131–151.
Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141.
Ryu, H. J., Adam, H., & Mitchell, M. (2017). InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1712.00193.
Silva, S., & Kenney, M. (2019). Algorithms, platforms, and ethnic bias. *Communications of the ACM, 62*(11), 37–39.
Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.
Suresh, H., Gong, J. J., & Guttag, J. V. (2018). Learning Tasks for Multitask Learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 802–810).
Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research, 51*(4), 463–479.
Ukanwa, K., & Rust, R. T. (2020). Discrimination in service. *Working paper*.
Vigdor, N. (2019). *Apple Card Investigated after Gender Discrimination Complaints*. Retrieved from https://www.nytimes.com/2019/11/10/business/Apple-credit-cardinvestigation.html. Accessed January 6, 2022.
Weissman, J. (2018). *Amazon created a hiring tool using A.I. it immediately started discriminating against women.* https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html.
Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining.* In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1). London, UK: Springer-Verlag.
Wang, X. S., Ryoo, J. H. J., Bendle, N., & Kopalle, P. K. (2021). The role of machine learning analytics and metrics in retailing research. *Journal of Retailing, 97*(4), 658–675.
Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science, 66*(3), 1045–1070.
Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. *International Conference on Machine Learning*, 325–333.
Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Benjamin van Giffen (Ph.D., University of St.Gallen) is an Assistant Professor at the University of St.Gallen, Switzerland. He is also founder and head of the Management of AI Lab at University of St.Gallen. His research and teaching focuses on the organizational adoption of artificial intelligence, AI business value, digital platforms and ecosystems, and human-centered design innovation (e.g., Design Thinking).

Dennis Herhausen (Ph.D., University of St.Gallen) is Associate Professor of Marketing at Vrije Universiteit Amsterdam. His research, teaching, and executive education revolve around the themes of digital communication, customer journeys and experience, multi-channel management, digital capabilities, and social media management. His work has been funded by national and international research grants, has received several awards, and is published in the *Journal of Marketing*, *Journal of Marketing Research*, *Journal of the Academy of Marketing Science*, and *Harvard Business Review*, among others.

Tobias Fahse is a Research Associate and PhD Candidate at the University of St.Gallen, Switzerland, and an associated researcher at the Management of AI Lab at University of St. Gallen. His research focuses on bias in machine learning, explainable AI, and AI business value, each with applications in ML-based demand forecasting.