Bus Inf Syst Eng 64(3):335–348 (2022)
https://doi.org/10.1007/s12599-021-00716-w

RESEARCH NOTES

The Cost of Fairness in AI: Evidence from E-Commerce

Moritz von Zahn · Stefan Feuerriegel · Niklas Kuehl

Received: 14 July 2020 / Accepted: 9 June 2021 / Published online: 7 September 2021

Abstract Contemporary information systems make widespread use of artificial intelligence (AI). While AI offers various benefits, it can also be subject to systematic errors, whereby people from certain groups (defined by gender, age, or other sensitive attributes) experience disparate outcomes. In many AI applications, disparate outcomes confront businesses and organizations with legal and reputational risks. To address these, technologies for so-called "AI fairness" have been developed, by which AI is adapted such that mathematical constraints for fairness are fulfilled. However, the financial costs of AI fairness are unclear. Therefore, the authors develop AI fairness for a real-world use case from e-commerce, where coupons are allocated according to clickstream sessions. In their setting, the authors find that AI fairness successfully manages to adhere to fairness requirements, while reducing the overall prediction performance only slightly. However, they find that AI fairness also results in an increase in financial cost. Thus, in this way the paper's findings contribute to designing information systems on the basis of AI fairness.

Keywords AI fairness · Algorithmic fairness · Fair AI · Costs · Artificial intelligence · Machine learning

Accepted after two revisions by Dennis Kundisch.

M. von Zahn (✉)
Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany
e-mail: vzahn@wiwi.uni-frankfurt.de

M. von Zahn · S. Feuerriegel
ETH Zurich, Weinbergstr. 56/58, 8092 Zurich, Switzerland

N. Kuehl
Karlsruhe Institute of Technology, Kaiserstraße 89, 76133 Karlsruhe, Germany

1 Introduction

Contemporary information systems make widespread use of artificial intelligence (AI). Artificial intelligence helps value creation (e. g., Müller et al. 2018), yet it is continuously confronted with ethical issues and fairness laws (Hacker 2018; White and Case 2017). For instance, AI can lead to disparate outcomes for people according to certain sociodemographics (gender, race, or other attributes deemed sensitive). In this case, AI[1] may lead to discrimination (Barocas and Selbst 2016).

Empirical evidence has confirmed disparate outcomes in a variety of AI use cases. In credit scoring, AI has been found to deny loan applications from women and racial minorities at a disproportionately high rate (Hardt et al. 2016). In the criminal justice system, AI is being increasingly utilized to predict the risk of recidivism, but it has falsely classified black defendants as "at risk" more frequently than non-black defendants (Angwin et al. 2016). In e-commerce, AI is utilized to personalize website interactions, and yet it has been found that AI systems show significantly fewer advertisements for high-paying jobs to women than to men (Datta et al. 2015; Lambrecht and Tucker 2019). This could limit women's access to resources or hinder economic advances.

In order to overcome fairness issues in AI, prior literature has developed algorithms for so-called "AI fairness"

[1] Many AI applications that are subject to fairness issues originate from the subdomain of supervised machine learning (including this study). Fairness is also a concern in other areas of AI such as unsupervised learning (Garg et al. 2018) and even rule-based inferences. Hence, we follow follow the terminology from Russell et al. (2015) and utilize the term "AI" consistently for any type of inference engine, as this allows us to highlight that the implications of discrimination in AI are of widespread applicability.

(cf. Feuerriegel et al. 2020; Haas 2019). AI fairness makes it possible to build inferences that satisfy mathematical definitions of fairness and that do not lead to disparate outcomes for certain individuals (Dwork et al. 2012; Hardt et al. 2016). Intuitively, it might seem sufficient to simply omit sensitive attributes. However, other attributes may serve as proxies and, as a result, the source of unfairness may persist (Barocas and Selbst 2016). This is best illustrated by means of an example. Salary may serve as a proxy for gender. Therefore, even if gender is removed, AI can leverage one of the proxies and thus lead to outcomes that discriminate by gender. A remedy is provided by AI fairness, which is designed so that certain mathematical constraints are fulfilled in the interest of fairness (see Sect. 2 for an introduction).

From a theoretical viewpoint, the use of AI fairness should have implications for the underlying prediction performance. This is because AI fairness introduces additional mathematical constraints and thus changes the parameter search space (Wick et al. 2019). This may eventually affect the underlying prediction performance and, therefore, also the financial costs when AI fairness is deployed in information systems (IS) practice. However, empirical evidence quantifying the financial costs of AI fairness is lacking.

We study the financial costs of AI fairness in e-commerce due to several reasons. First, a lack of fairness in e-commerce might be unethical as it can have negative implications for users (Susser et al. 2019). For instance, targeted advertising based on AI has been found to be biased by gender (Lambrecht and Tucker 2019), where, as a result, women are withheld from seeing ads related to high-paid jobs. Second, a lack of fairness in AI for e-commerce may be unlawful. This is best explained through an example where an AI application awards users digital coupons with discounts and thus incentivizes a purchase (Koehn et al. 2020). Here AI may lead to a coupon distribution according to which users of certain sociodemographics are favored at a disproportionate rate. This could violate fairness laws, as previous research has argued (Hacker 2018; White and Case 2017; Barocas et al. 2019). Third, a lack of fairness in e-commerce carries reputational risks. A prominent example from the United States concerns the retailer Staples, which leveraged analytics to offer discounts based on geographic properties. Later, it was found that discounts were unevenly distributed and primarily targeted neighborhoods with high-income households (while withholding discounts from low-income households). This was perceived as "highly discriminatory" by users (Valentino-Devries et al. 2012). Such reputational risk can be mitigated by IS practitioners through the use of AI fairness.

In this paper, we implement AI fairness in a use case from e-commerce where the retailer aims at steering users towards purchase by allocating personalized digital coupons. A coupon is issued if a user is at risk of exiting the e-commerce website with no purchase, which we predict based on real-world clickstream data from a large online retailer. For the retailer, interventions through digital coupons incur financial costs in the form of lost profits (e. g., if a coupon was not issued and where, as a result, the user left the website without generating profits from a purchase). For the underlying predictions, we implement AI fairness that treats gender as a sensitive attribute. We then compare our implementation of AI fairness against those obtained from a default application of AI without considering fairness constraints. Based on this, we quantify the financial costs of AI fairness.

2 Background

2.1 Fairness in AI

Fairness in AI is mathematically formalized through so-called fairness notions, which measure deviations from an outcome that would be regarded as fair (Chouldechova and Roth 2020). However, different notions exist, and it is mathematically impossible to fulfill all notions of fairness at the same time (Kleinberg et al. 2016). Therefore, IS practitioners need to choose a fairness notion that is appropriate to the given use case. For a detailed overview of fairness notions, we refer to Barocas et al. (2019). In the following, we provide a brief summary of two fairness notions – i. e., statistical parity and equalized odds – that are particularly relevant to IS practice. For this, we use the following notation: we refer to the predicted label as $\hat{Y}$, the actual label as Y, and the sensitive attribute as A.

Statistical parity (also called demographic parity and equal parity) requires that the predicted label $\hat{Y}$ is independent of the sensitive attribute A (Dwork et al. 2012). In other words, the likelihood of outcomes should be the same across the protected group (e. g., female users) and outside of it. Formally, this is given by

$$P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y} \,|\, A = a), \quad \forall a \in A. \tag{1}$$

For instance, in e-commerce, statistical parity would require users selected by AI to receive digital coupons to reflect an equal distribution of male and female users. This definition of fairness is common in real-world applications and many legal frameworks (Feldman et al. 2015; Barocas and Selbst 2016). Due to its legal relevance, statistical parity is used in our empirical study in e-commerce. However, inherent to statistical parity is that it only

considers the predicted label, but neither the actual distribution of labels nor the error rates when making inferences. As such, statistical parity ignores any possible correlation between Y and A, which is often not desirable in cases of different base rates (e. g., because the website has primarily customers from one gender).

Equalized odds refers to independence between the sensitive attribute A and both type-I/type-II errors (Hardt et al. 2016). This notion is especially useful for cases in which a positive prediction provides a specific benefit, yet where errors in granting this benefit should be equal within the protected group and outside of it. Formally, inferences satisfy the notion of equalized odds with respect to a sensitive group if both outcome $\hat{Y}$ and A are independent conditional on the actual distribution of labels Y. This is given by

$$P(\hat{Y} = \hat{y} \mid Y = y) = P(\hat{Y} = \hat{y} \mid Y = y, A = a), \quad \forall a \in A, \quad \forall y \in Y. \tag{2}$$

When applied to our previous example from e-commerce, equalized odds suggests that (1) the probability of users eligible for a coupon being identified as such must be the same for male and female users, and (2) the probability of users that are not eligible to still receive a coupon must also be the same for male and female users. Hence, equalized odds as a fairness notion is beneficial for IS use cases where disparities in the actual distribution of labels Y can be considered legitimate (and not unfair).

In our e-commerce study, we consider both statistical parity and equalized odds. We provide further details on how we measure the level of fairness as part of our empirical setting (Sect. 3.6).

2.2 Implementations of AI fairness

Different algorithms have been developed for implementing AI fairness (Holstein et al. 2019). In general, these algorithms target a specific notion of fairness and are typically designed to mitigate between-group disparities in the predictions. For this, AI fairness explicitly requires access to the sensitive attribute at the time of training.

A naïve strategy for achieving fairness may be to simply omit the sensitive attribute, which we refer to as "blinding". However, this has been found to be insufficient for achieving fairness as other attributes may carry information pertaining to the sensitive attribute and thereby serve as proxies (e. g., Barocas et al. 2019). For instance, in e-commerce, AI might consider a user's browser history to determine advertising content. However, information on browser history is a proxy for gender (e. g., github.com as a proxy for men and pinterest.com as a proxy for women; Barocas et al. 2019). Instead, one needs to leverage algorithms from AI fairness.

Algorithms for AI fairness can be grouped according to the stages at which fairness enters the AI pipeline (Barocas et al. 2019). First, preprocessing algorithms transform the underlying data so that potential biases are mitigated (e. g., reweighing; Kamiran and Calders 2012). Second, in-processing algorithms change the underlying classifier so that fairness becomes part of the objective (e. g., adversarial debiasing; Zhang et al. 2018). Third, post-processing algorithms adjust the predictions post hoc (e. g., reject option based classification; Kamiran et al. 2012). If fairness enters the AI pipeline at an early stage, it might be reverted again at a later stage. In order to circumvent this issue, we primarily rely on post-processing in the form of reject option based classification.

2.3 Cost-Fairness Tradeoff

Prior research has studied AI fairness in terms of its tradeoff between fairness and prediction performance. Corbett-Davies et al. (2017) inferred the tradeoff between prediction performance and fairness when predicting the risk of recidivism. Friedler et al. (2018) compared the tradeoff across different algorithms for AI fairness. A similar approach was taken by Haas (2019), who presented a framework that can introduce different levels of fairness in order to balance fairness against prediction performance. The aforementioned studies often use the term "costs" to refer to the loss function measuring the prediction performance, whereas the actual *financial* costs have been overlooked. Hence, to the best of our knowledge, we offer the first research to examine an economic impact of AI fairness.

The use of AI fairness might theoretically impact financial costs in either a positive or negative direction. On the one hand, AI fairness typically improves the prediction performance for the protected group (e. g., Hardt et al. 2016) but lowers it for the non-protected group, which often represents the larger user base. Hence, one could expect that this results in larger overall financial costs. On the other hand, if prediction errors for the protected group are comparatively expensive, fewer prediction errors in the protected group might decrease overall financial costs. Hence, the overall relationship between prediction performance and financial costs in AI fairness is complicated and motivates our research question:

> *How does AI fairness for price promotions in e-commerce affect financial costs?*

In the following, we empirically evaluate this research question based on a real-world use case from e-commerce.

3 Empirical Setting

3.1 AI fairness for E-Commerce

In e-commerce, retailers aim at steering user behavior towards purchasing and, for this purpose, make use of AI. In particular, AI can be utilized by online retailers to target users exiting their website with no purchase. By predicting whether a user will exit with no purchase, online retailers can trigger personalized interventions (e. g., digital coupons) to steer users towards making a purchase (Gofman et al. 2009; McDowell et al. 2016; Ding et al. 2015).

Predictions in e-commerce typically build upon clickstream data. Clickstream data records the behavior of users on a website. It comprises information such as the pages visited, the time spent on each page, and the overall number of user interactions in the form of clicks. Clickstream data can be leveraged by AI to predict the risk that a user will exit with no purchase (Montgomery et al. 2004; Hatt and Feuerriegel 2020). These predictions have been built upon linear models (Olbrich and Holsing 2011), neural networks (Jenkins 2019; Sheil et al. 2018), or boosting (de Bock and van den Poel 2010), often in combination with feature engineering in order to accommodate the sequential structure of clickstream data (e. g., Baumann et al. 2019). For a detailed overview of clickstream analytics, we refer to Mobasher (2007).

Clickstream analytics may yield disparate outcomes with regard to gender. If users from one gender exhibit different clickstream behavior, then this is likely to be reflected in the predictions based on clickstream data. Hence, users from that gender might be – depending on the intervention – favored or disadvantaged in a disproportionate manner. This is best illustrated by considering an example. In general, users benefit from digital coupons through reduced prices (Reimers and Xie 2019). However, groups of users who systematically produce a lower rate of mouse clicks receive more coupons and, hence, are disproportionately favored. Notably, it has been proven that gender biases are present in real-world clickstream settings. For instance, it was found that clickstream data can indeed be utilized in order to predict the gender of users (de Bock and van den Poel 2010) and that women see significantly fewer online ads related to high-paying jobs than men (Datta et al. 2015; Lambrecht and Tucker 2019). Online ads and digital coupons may differ in the type of benefit they provide, still both illustrate how AI in e-commerce can lead to disparate outcomes. Needless to say, if online retailers are interested in remedying such disparate outcomes, they could implement AI fairness as shown in the following.

In our evaluations, the fairness notion is first set to statistical parity due to its widespread use in the legal domain (Feldman et al. 2015). We then expand the study to include equalized odds, which shifts the focus towards prediction performance responsible for offering coupons.

3.2 Data Description

We evaluate AI fairness based on real-world clickstream data from *Digitec Galaxus*. Digitec Galaxus is the largest online retailer in Switzerland, offering more than a million different products with an emphasis on consumer electronics. The company's website offers a diverse range of information, including product reviews.

Our data consists of the complete set of clickstream sessions that we collected over the course of one week in the summer of 2019. Each session is of variable length and corresponds to the sequence of pages visited. Furthermore, for each page in that sequence, the following three variables were retrieved: (1) the visit depth, that is the number of pages visited before the given page; (2) the time spent on the given page; (3) the number of visits to pages within the category to which the current page belongs. In addition, for every session, the data comprises the total number of pages visited, the total duration, and the age and gender of the user. Finally, for each session, the prediction label denotes whether a purchase or an exit with no purchase took place. The data was preprocessed in a manner analogous to that found in prior literature (e. g., Montgomery et al. 2004), which is detailed in "Appendix 1" (available online via http://link.springer.com).

In our e-commerce setting, the objective is to identify users at risk of exiting with no purchase and, once identified, to steer them towards purchase by providing a coupon. Hence, this yields different financial costs for type-I/type-II errors in the prediction. To account for this, financial costs (here: lost profits) were assigned to different prediction errors as would arise for such a user exit prediction; see Table 1. For reasons of confidentiality, these costs were calculated based on industry-wide operating margins. A false positive represents a type-I error whereby the exit was falsely predicted and thus incurred costs for an unnecessary coupon. A false negative represents a type-II error whereby the purchase was falsely predicted and, due to the absence of a coupon, resulted in the loss of a potential sale. Lost

Table 1 Financial costs (lost profits) of prediction errors by gender

	False negative ($\hat{Y}$ = "purchase", Y = "no purchase")	False positive ($\hat{Y}$ = "no purchase", Y = "purchase")
Female	USD 2.82	USD 2.50
Male	USD 4.24	USD 2.50

sales are discounted, as coupons do not always have the desired effect on each user and, if they do, reduce the sales price itself. Therefore, based on the operating margin, the discounted average profit per purchase is assigned to the type-II error. We distinguish the average profit by gender, as the average sales volume also differs by gender.

3.3 Descriptive Statistics

The preprocessed data sample comprises 400 clickstream sessions. The descriptive statistics of the sessions are reported in Table 2. All variables are transformed to preserve confidentiality (i. e., this maintains the relative distribution but units are sanitized). Overall, clickstream sessions reveal pronounced differences between female and male users. On average, female users browse more pages per session, spend more time on each page, and visit the same page category more frequently. This is also reflected in the different quantiles of the summary statistics. Moreover, user age is distributed differently for each gender, with female users corresponding to a higher mean age, but a lower median age, than male users. Hence, in our data, several attributes relay information on gender and may serve as proxies.

The data is highly imbalanced with regard to both gender (four-fifth are men) and the prediction label (most of the sessions are exited with no purchase). For comparison, industry averages have estimated the ratio of sessions with no purchase to 97 % (Statista 2020). The ratio of user sessions with no purchase reveals a gender imbalance, i. e., it is larger for male than for female users. Hence, not only the clickstream data, but also the purchase behavior itself is subject to gender differences (Digitec Galaxus 2018).

3.4 Prediction Framework

We estimate different classifiers – namely (1) a default, (2) a blinded, and (3) a fair classifier – for the purpose of predicting user exits with no purchase:

(1) *Default classifier* This approach represents the status quo that is currently utilized in clickstream analytics (cf. Baumann et al. 2019). The default classifier has access to the sensitive attribute and, on top of that, its inferences are not bound by fairness constraints. In our analysis, we report results for when the classifier is implemented by extreme gradient boosting (e. g., as in Senoner et al. 2021). Other classifiers are part of the robustness checks. For all, the classification threshold is chosen based on the training set so that the financial costs of prediction errors are minimal.

(2) *Blinded classifier* This classifier is analogous to the default classifier and makes predictions without constraints for ensuring fairness. Yet it differs from the default classifier in one regard: the sensitive attribute is omitted. Nevertheless, it may still infer information concerning gender from other features that act as proxies. For instance, as mentioned above, female users might be characterized by different clickstream behavior whereby they spend more time on pages than male users (which is supported by our descriptive statistics). Again, the classifier is implemented via boosting.

(3) *Fair classifier* This classifier makes predictions (based on boosting, as above) while explicitly accommodating fairness constraints, i. e., statistical parity or equalized odds with regard to gender as the sensitive attribute. In our work, we implement fairness via reject option based classification as a post-processing technique (Kamiran et al. 2012). Post-processing techniques are well suited to real-world settings. In contrast to preprocessing, they introduce fairness at a later stage in the prediction pipeline. This is especially beneficial for imbalanced datasets where fairness from preprocessing is reverted by a classifier as a means of reweighing samples to counteract the imbalances. In addition, unlike in-processing, post-processing techniques are flexible in terms of the underlying classifier.

Table 2 Descriptive statistics by gender

Variable	Mean		Standard deviation		25 % quantile		Median		75 % quantile	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Gender	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
Age	1.662	1.608	0.458	0.508	1.300	1.257	1.539	1.560	1.950	1.950
Visit depth	0.457	0.409	0.289	0.268	0.211	0.211	0.386	0.331	0.622	0.579
Duration per page	0.151	0.145	0.241	0.259	0.004	0.000	0.067	0.046	0.177	0.162
Cumulative number of clicks	0.394	0.333	0.311	0.272	0.132	0.132	0.353	0.249	0.534	0.448

Reject option based classification achieves fairness by altering the prediction of uncertain instances within a confidence band (Kamiran et al. 2012). This confidence band is of variable length and centered around the classification threshold (e. g., if the threshold is 0.5, the interval of [0.4, 0.6] may be a suitable confidence band). The width of this band is determined based on training data and depends on the magnitude of the disparities in the predictions, the level of fairness to be achieved, and the corresponding financial costs. Specifically, among all different widths achieving the required level of fairness, the one yielding the lowest financial cost is chosen. Within the confidence band, instances are flipped, which means that the favorable labels (i. e., exit with no purchase) are replaced by non-favorable labels (i. e., purchase) and vice versa. This is done based on group membership, i. e., based on gender. Specifically, an instance is flipped if two conditions are met: first, the score of the prediction must be within the confidence band and, second, the instance must either hold the non-favorable label and be of the protected group (female) or hold the favorable label outside the protected group (male). This results in a higher number of favorable labels within the protected group and a lower number outside of it, which may also imply a shift in prediction errors. The algorithm thereby mitigates between-group disparities. Our implementation is based on the "aif360" library.[2]

Later, we perform analyses with two additional classifiers, namely a lasso and a deep neural network. The lasso performs an implicit variable selection in order to avoid overfitting. Deep neural networks are known for being highly flexible as they can model complex non-linearities, and yet their advantages often become evident only in applications with large-scale datasets (Kraus et al. 2020).

3.5 Estimation Details

The data was randomly split into different sets for training and testing, following common conventions (Hastie et al. 2017). Formally, we ensure an equal proportion of labels within the sets by performing stratified sampling. All classifiers utilize the same 80 % of data for training and the same 20 % for testing. For the fair classifier, 20 % of the training set is reserved as a validation set based on which the parameters of the fairness algorithm are determined (i. e., the confidence band of reject option based classification). Analogous to prior literature on AI fairness (Friedler et al. 2018), we perform a total of 30 random train-test splits, that is, we repeat all computational experiments across 30 runs. All results are reported for the test set and thus for out-of-sample data.

The classifiers are trained as follows. The hyperparameters for boosting are determined by applying a grid search in a 5-fold cross-validation on the training set (tuning ranges are reported in "Appendix 2"). All robustness checks with the lasso and the neural network are implemented accordingly.

3.6 Performance Metrics

Different performance metrics are utilized. We draw upon (1) the overall prediction performance, (2) fairness metrics, and (3) financial costs, as detailed in the following.

The prediction performance in detecting users at risk of exiting with no purchase is based on the area under the receiver operator characteristic, or AUROC for short. The AUROC accounts for imbalances in the dataset. For comparison, we also report the balanced accuracy, the F1-score, and the AUPRC (area under the precision recall curve).

Fairness is quantified according to the notions of statistical parity (Dwork et al. 2012) and equalized odds (Hardt et al. 2016). Statistical parity is selected as our default metric for assessing fairness due to its widespread use in the legal domain (Barocas and Selbst 2016). Formally, one computes the difference in the probability of labeling female vs. male users as being about to exit with no purchase and thus receiving a coupon. Let us consider an example in which the proportion of female users that are predicted to exit with no purchase is 0.75, and the proportion of male users anticipated to do so is 0.80. In this example, the difference amounts to $0.75 - 0.80 = -0.05$; that is, the probability of receiving a coupon is 5 percentage points lower for female than for male users. For comparison, a value of 0 is considered fair. For a given classifier, the level of fairness is computed using

$$\text{STATISTICAL PARITY} = P(\hat{Y} = \text{``no purchase''} \mid \text{``female''}) - P(\hat{Y} = \text{``no purchase''} \mid \text{``male''}). \quad (3)$$

Equalized odds measures the difference in the error rates between male and female users (Hardt et al. 2016). For certain products in e-commerce, equalized odds can be a useful alternative to statistical parity, especially when one gender is represented more frequently among the buyers. For instance, for high heels, it may not make sense to aim

[2] IBM AI Fairness 360 Open Source Toolkit (aif360): https://aif360.mybluemix.net/.

at fairness as defined by statistical parity, which would aspire to an equal probability of coupons for female vs. male users. Due to the nature of the product, the proportion of potential buyers is higher among female than male users. Therefore, equalized odds might be the preferred notion of fairness, which would simply ensure that a male user interested in high heels has the same chance of being selected for a coupon as an interested female user. In our setting, male users who ultimately exit with no purchase may have a probability of 0.05 of being classified incorrectly. For female users, let the probability amount to 0.10, corresponding to a difference of −0.05 in the false negative rate. As a consequence, the female users who intend to exit without purchase are 5 percentage points less likely to benefit from coupons than like-minded male users. Formally, the false positive rate *FPR* refers to the probability that users who intend to purchase a product are being classified incorrectly. Both *FNR* and *FPR* are combined when computing the level of fairness according to equalized odds via

$$\text{EQUALIZED ODDS} = \frac{1}{2}\left[(FPR_{\text{female}} - FPR_{\text{male}}) + (FNR_{\text{male}} - FNR_{\text{female}})\right], \tag{4}$$

where FPR_{female} refers to the false positive rate among female users and where FPR_{male}, FNR_{female}, and FNR_{male} are defined analogously.

Financial costs are computed by weighting the confusion matrices of female and male users with the corresponding costs in USD (given in Table 1). Hence, this takes into account differences between the financial costs associated with type-I and type-II errors. Formally, we add up the costs caused by errors made on the test data and then divide the total costs by the number of samples. This is given by

$$\text{FINANCIAL COSTS} = \frac{2.50\,(FP_{\text{female}} + FP_{\text{male}}) + 2.82\,FN_{\text{female}} + 4.24\,FN_{\text{male}}}{TP + TN + FP + FN}, \tag{5}$$

where *FP* refers to the total number of false positives, FP_{female} to the number of false positives for female users, and all other variables are defined analogously.

4 Empirical Results

The results are reported for both statistical parity (Sect. 4.1) and equalized odds (Sect. 4.2). We focus on the results generated on the basis of boosting, as this model achieved the best overall prediction performance and is thus favored during model selection. Afterwards, we replace boosting with other classifiers (Sect. 4.3). Finally, we provide additional analysis to interpret our results (Sect. 4.4).

4.1 Results for Statistical Parity

The results are reported in Table 3 (panel: boosting). The default and blinded classifier yield a similar prediction performance, whereas the prediction performance of the fair classifier is slightly inferior. However, the fair classifier yields a significantly higher level of fairness (i. e., as defined by statistical parity) than the default and blinded classifiers. The default classifier results in a statistical parity metric of −0.077. This means that the probability of receiving a coupon is 7.7 percentage points higher for male than for female users. The blinded classifier performs slightly better, with the probability being 3.1 percentage points higher for male users. The fair classifier achieves nearly full statistical parity, with the probability of receiving a coupon being 0.1 percentage points lower for male users.

Introducing AI fairness results in higher financial costs. Specifically, the fair classifier results in costs of USD 0.551 as compared to USD 0.508 for the default classifier. This corresponds to an increase of 8.5 %. In sum, replacing the default with the fair classifier leaves the prediction performance only slightly diminished, but improves fairness at a higher financial cost.

4.2 Results for Equalized Odds

The results for equalized odds are reported in Table 4 (panel: boosting). Similar to the results for statistical parity, the fair classifier yields a slightly lower prediction performance than both the default and blinded classifiers. Furthermore, the fair classifier also yields a higher level of fairness (i. e., as defined by equalized odds) than the other classifiers, reducing the equalized odds metric to −0.028 from −0.127 and −0.067, respectively. In our setting, this

Table 3 Performance metrics for statistical parity

Prediction model	Classifier	Balanced accuracy	F1-score	AUROC	AUPRC	Fairness metric (statistical parity)	Financial costs (in USD)
Boosting	Default	0.649	0.881	0.727	0.855	− 0.077	0.508
	Blinded	0.642	0.879	0.725	0.852	− 0.031	0.525
	Fair	0.644	0.875	0.705	0.836	0.001	0.551
Lasso	Default	0.521	0.847	0.596	0.790	− 0.040	0.672
	Blinded	0.522	0.849	0.600	0.793	− 0.027	0.663
	Fair	0.517	0.845	0.519	0.729	0.007	0.683
Neural network	Default	0.651	0.883	0.664	0.808	− 0.048	0.498
	Blinded	0.655	0.886	0.670	0.807	− 0.009	0.485
	Fair	0.647	0.874	0.679	0.825	0.014	0.554

Stated: mean value over 30 random train-test splits

Table 4 Performance metrics for equalized odds

Prediction model	Classifier	Balanced accuracy	F1-score	AUROC	AUPRC	Fairness metric (equalized odds)	Financial costs (in USD)
Boosting	Default	0.649	0.881	0.727	0.855	− 0.127	0.508
	Blinded	0.642	0.879	0.725	0.852	− 0.067	0.525
	Fair	0.640	0.874	0.707	0.846	− 0.028	0.554
Lasso	Default	0.521	0.847	0.596	0.790	− 0.052	0.672
	Blinded	0.522	0.849	0.600	0.793	− 0.043	0.663
	Fair	0.517	0.846	0.518	0.731	− 0.009	0.679
Neural network	Default	0.651	0.883	0.664	0.808	− 0.060	0.498
	Blinded	0.655	0.886	0.670	0.807	− 0.021	0.485
	Fair	0.651	0.877	0.610	0.771	0.014	0.536

Stated: mean value over 30 random train-test splits

means that the difference between the prediction performance for male and female users declines by 9.9 and 3.9 percentage points, respectively. Hence, introducing the fair classifier promotes equal chances of being financially (dis)favored between women and men.

The financial impact of AI fairness is as follows. The fair classifier results in higher costs than the default classifier (USD 0.554 as compared to USD 0.508). This corresponds to an increase of 9.1 %, which is of similar magnitude as the result for statistical parity.

4.3 Sensitivity Analysis

In addition to boosting as the underlying prediction algorithm, we have repeated all experiments based on the lasso and a deep neural network. For both, we yield results that mainly support our initial findings. In addition to reject option based classification as the underlying algorithm for AI fairness, we have repeated the experiments with a preprocessing algorithm (i. e., reweighing; Kamiran and Calders 2012) and an in-processing algorithm (i. e., adversarial debiasing; Zhang et al. 2018). However, we found both algorithms to be ineffective in our setting. More details on the sensitivity analysis can be found in "Appendix 3".

4.4 Interpretation

We find two aspects of our results particularly noteworthy. Therefore, we provide interpretations for (i) the level of fairness provided by the blinded classifier and (ii) the differences in prediction performance and financial costs between the fair and default classifiers.

The blinded classifier provides a level of fairness that lies between the levels of the other classifiers. The reason is that the disparities can have two different causes: First, disparities can be induced directly by the sensitive attribute. However, sensitive attributes are only present in the default classifier, while they are omitted in the blinded classifier. Specifically, the sensitive attribute "gender" is leveraged by the default classifier to predict whether a user will exit with no purchase, particularly for female users. This is supported when analyzing the underlying variable importance (Fig. 1). As a result, female users yield

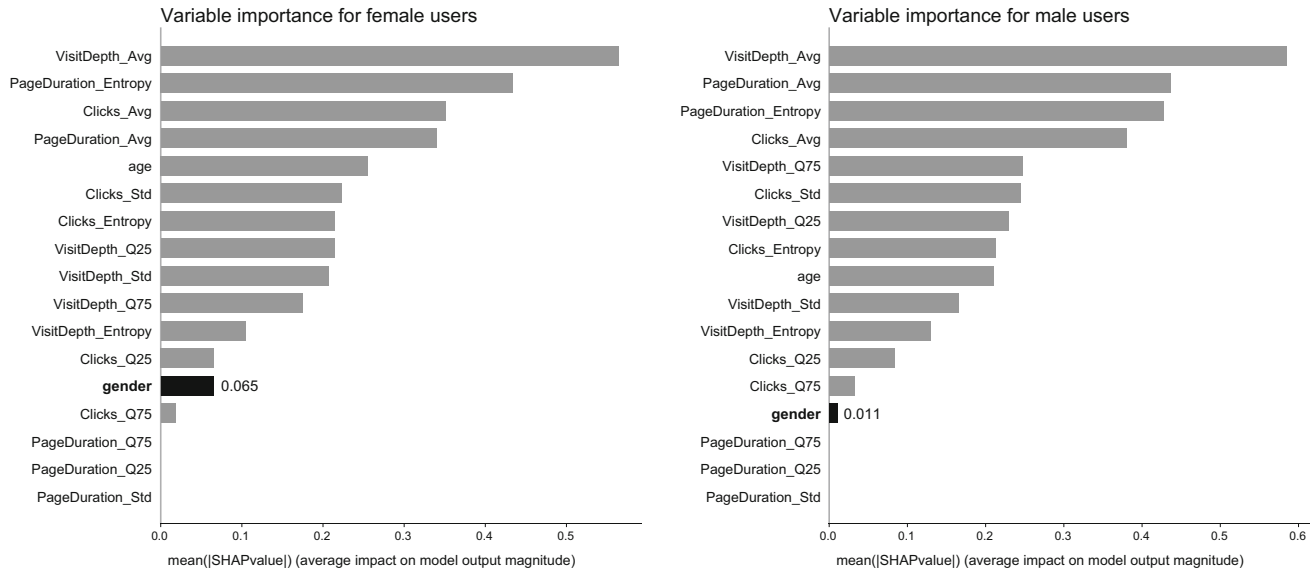

Fig. 1 The variable importance for the predictions of the default classifier based on boosting are shown for female (left) and male users (right). On the vertical axis, the clickstream and user attributes are listed. The horizontal axis shows the absolute average SHAP value, indicating the impact of the attribute on the prediction (Lundberg and Lee 2017)

disparate outcomes, i. e., fewer coupons, in the case of the default classifier. However, the sensitive attribute is absent in the blinded classifier. Second, disparities are induced by proxies. For instance, the average visit depth is both a proxy for gender (Fig. 2) and further appears to be an important predictor (Fig. 1). As a result, the blinded classifier is subject to the disparities from proxies, but not to disparities induced directly by the sensitive attribute. In contrast, the default classifier is subject to both, while the fair classifier mitigates both.

When comparing the fair classifier to the default classifier, we observe significantly higher financial costs despite an only slightly lower prediction performance. This is due to imbalances in the cost structure: costs are differently distributed (by gender and classification outcome) than prediction errors (by classification outcome only). Specifically, Table 5 shows that for female users, the default classifier triggers coupons for 83.7 % of users and yields an accuracy of 0.833. After introducing fairness, i. e., the fair classifier with statistical parity, 89.1 % of female users receive a coupon with the respective accuracy decreasing to 0.809.[3] Here the number of false negatives decreases, but the number of false positives increases more sharply. For male users, the default classifier triggers coupons for 91.5 % of users and yields an accuracy of 0.799, whereas the fair classifier triggers coupons for 88.9 % of users and yields a similar accuracy of 0.793. In this case, the number of false positives decreases to almost the same degree as the number of false negatives increases. Hence, the accuracy for male users is barely affected by introducing the fair classifier. However, the financial costs associated with false negatives for male users are particularly high (USD 4.24, see Table 1) due to the high expected sales volume for male users for a true positive, i. e., if a coupon had been triggered. As a consequence, a large extent of the overall cost increase (i. e., 76.7 %) is driven by the shift in prediction errors for male users.

In sum, we find that the fair classifier building upon reject option based classification consistently provides a high level of fairness. In addition, for the two classifiers with the highest prediction performance (boosting algorithm and deep neural network), the fair variant results in higher financial costs than the default classifier.[4] This increase is primarily due to AI fairness shifting the distribution of errors towards more expensive prediction errors. Specifically, the financial costs increase by 8 to 10 %.

5 Discussion

Prior research has found that AI fairness typically lowers the prediction performance (e. g., Kamiran and Calders 2012), but its financial costs have remained unclear. As

[3] Notably, the lower prediction performance for female users is different form findings in, e. g., Hardt et al. (2016), where fairness increases the prediction performance for the protected group.

[4] We further compare the results to a cost-optimal classifier, which would give an lower bound to the price of AI fairness. For this, we draw upon a perfect (error-free) classifier and then compute the corresponding costs due to ensuring the fairness constraint from statistical parity (i. e., by providing additional coupons to female users until the share is equal to that of male users). In this hypothetical setting, statistical parity would inflict costs of USD 0.021.

Table 5 Confusion matrices of the default, blinded, and fair classifier for both female and male users

	Default classifier				Blinded classifier				Fair classifier			
	Female		Male		Female		Male		Female		Male	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	10.80	2.20	46.90	12.07	11.03	2.60	46.60	11.97	11.03	2.80	45.90	11.43
$\hat{Y} = 0$	0.40	2.13	0.90	4.60	0.17	1.73	1.20	4.70	0.17	1.53	1.90	5.23

$\hat{Y} = 0$ refers to a predicted purchase (no coupon provided) and $\hat{Y} = 1$ refers to a predicted exit without purchase (coupon provided). The results rely on boosting for the underlying prediction algorithm and on statistical parity as the notion of fairness

Stated: mean value over 30 random train-test splits

detailed above, the relationship between a lower overall prediction performance and financial costs is non-trivial and requires a thorough empirical analysis. Hence, we expand over the body of knowledge by providing real-world estimates on the financial costs of AI fairness in a use case from e-commerce. In the following, we will discuss the implications of our work (Sect. 5.1) and future opportunities for IS research (Sect. 5.2).

5.1 Implications

Our findings entail several implications. For academics, we provide real-world estimates of the actual financial costs due to AI fairness, which is an insight frequently requested by the research community (e. g., Chouldechova and Roth 2020; Smith and Neupane 2018). Thereby, we support the further development of AI-based information systems and their degree of maturity within socio-technical systems (Maedche et al. 2019).

For managers and practitioners, we provide multiple relevant insights. First, we raise awareness of the fact that the introduction of AI fairness may have negative financial impacts. While we only analyzed one specific case and the results are not generalizable, we demonstrate that introducing AI fairness was associated with financial costs in our e-commerce example. It could also lead to unethical outcomes whereby some individuals are left in a more inferior position – as they received fewer coupons. However, there might also be cases where AI fairness has no financial cost. For our e-commerce setting, we further find that the financial costs due to statistical parity and equalized odds are of similar magnitude. This is important as it means that costs might not be an argument when companies choose upon a fairness notion.

In terms of design choices for AI fairness, we see the following approaches deducted from our e-commerce application. As shown above, one approach demonstrated within this work is related to the "fairness by design" paradigm for AI (Abbasi et al. 2018), i. e., implementing post-processing techniques for constant fairness correction. For our default classifier, we observe that the difference in statistical parity is 7.7 %. This raises the question if full statistical parity is necessary. Some companies are likely to implement a relaxed version of statistical parity in which a certain difference in statistical parity is deemed tolerable. For instance, one such such relaxed variant is the 80 % rule.[5] If a company adopts the 80 % rule as a measure of

[5] The 80 % rule is common in many legal frameworks (Feldman et al. 2015; Barocas and Selbst 2016) and can be seen as a relaxed variant of statistical parity. The 80 % rule requires the share of favorable labels in the protected group to be at least 80 % of the share outside of it. This is fulfilled in our default classification, where the rate of coupons for female users is 92.3 % of that for male users.

fairness, our e-commerce use case would not require AI fairness. Hence, besides implementing AI fairness, another approach for companies might instead be to monitor the level of fairness and intervene only if a certain threshold is violated. Notwithstanding, IS practitioners might want to engage with AI fairness for a variety of reasons. In addition to legal and ethical considerations, unfairness in e-commerce might pose substantial reputational risks, as shown in the case of Staples (Valentino-Devries et al. 2012). Such reputational risks can be mitigated with state-of-the-art methods from AI fairness, such as the ones demonstrated in the present work.

Policymakers should be aware of the (potential) costs associated with AI fairness. By providing rigorous insights into the financial costs, we wish to stimulate broader societal discourse regarding AI fairness and particularly hope to raise awareness of its economic consequences (e. g., what price are we as society willing to accept for fair outcomes?). As our example shows, the introduction of fairness can link to additional costs for businesses, and increased opportunities for one group has adverse effects on multiple other stakeholders. Hence, fairness will not necessarily generate beneficial outcomes per se, but may lead to detrimental outcomes for some stakeholders. In fact, fairness may cause an overall negative effect. In the context of e-commerce, this will depend on the relative costs of a firms' potential adjustments, as this study shows. Within the fairness community, this is an often-discussed issue. Different notions of fairness contradict each other, and privileging one group may lead to the discrimination of another (Kleinberg et al. 2017). Therefore, policymakers should take both – potential financial costs and contradicting fairness perspectives – into account when drafting or amending legislation. For instance, the GDPR of the EU requires businesses to "ensure fair and transparent [data] processing" drawing upon "appropriate mathematical or statistical procedures" (GDPR 2016, Article 22, Recital 71). When enforcing these guidelines on national or state level, these mentioned tradeoffs should be kept in mind.

5.2 Opportunities for Future IS Research

Our empirical findings create many opportunities for future IS research. Most importantly, our findings are based on a particular empirical setting in e-commerce, namely clickstream analytics for price promotions. Future research should focus on exploring the financial costs of AI fairness in additional settings, i. e., how the 8 to 10 % cost increase would carry over to other use cases within, but also outside, the area of e-commerce. From a practical standpoint, it would be interesting to advance algorithms for AI fairness that can effectively deal with class imbalances, as these are widespread in real-world settings. It would be especially interesting to identify how class imbalances moderate the outcome of AI fairness. For this, our findings can provide researchers with a springboard (cf. Veale and Binns 2017).

A major aspect when dealing with AI fairness in IS is the choice of a suitable notion of fairness. As it is mathematically impossible for all notations of fairness to be fulfilled at the same time, the choice needs to be weighted carefully. As a first step, we started by investigating group fairness in the present work, namely statistical parity and equalized odds. Statistical parity in particular entails notable drawbacks in IS settings, as illustrated by our example of handing out coupons for high heels to female and male users equally. Hence, future research could add to the knowledge base by investigating other fairness notions and link them to user perceptions of what is considered fair. This is relevant as fairness ultimately represents a socio-technical construct and, hence, the definition of what is regarded as fair or unfair is not intrinsic to algorithms but rather lies in the hands of the programmers. Here the IS discipline is well suited to making relevant and impactful contributions.

This work studies the costs of AI fairness. If companies additionally gain a better understanding of the *value* of AI fairness, an economic weighing of both expenses and value added will allow managers to make better-informed decisions. Moreover, for a thorough economic analysis of AI fairness, further investigation is required into its effects on consumers, welfare, and social efficiency including the long-term returns. This is still an under-researched topic and thus represents an opportunity for the IS community to make a distinctive contribution.

6 Conclusion

Nowadays, information systems make widespread use of AI. However, AI might introduce disparate outcomes for users depending on certain sociodemographics, such as gender. A remedy to disparate outcomes has been developed in the form of AI fairness. In this work, we have quantified the financial costs of AI fairness based on an e-commerce application. We find that our fair classifier mitigates disparate outcomes, yet it also increases the financial costs by approximately 8 to 10 %. Thereby, our work represents an important empirical contribution for both research and IS practice.

Appendix 1: Data Preprocessing

The clickstream data was preprocessed in a manner analogous to that found in prior literature (e. g., Montgomery et al. 2004). First, we only extracted the pages that were

Table 6 Grid search for hyperparameter tuning

Model	Tuning parameter	Tuning range
Boosting	Minimum sum of weights in child	1, 2, 3, 4, 5, 6
	Maximum depth of trees	2, 3, 4, 5, 6, 7, 8
	Maximum delta steps	0, 1, 2, 3, 4
Lasso	Regularization strength α	10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}
Neural network	Learning rate	0.01, 0.05, 0.1
	Dropout rate	0.25, 0.5, 0.75
	Number of neurons	10, 12, 14, 16, 18, 20, 22
	Batch size	2, 4, 8, 16, 32

actually visited, that is, rendered in the browser window. Second, we omitted sessions that originated from web crawlers (this is based on a classification from the online retailer). Third, we assigned every page to one of the following categories: home, account, overview, product, marketing, content, community, and checkout (Montgomery et al. 2004). Fourth, we also filtered out the sessions containing either fewer than three pages or more than 50 pages visited. Fifth, we considered a session closed if the same page was open for longer than 20 minutes. Sixth, we only considered sessions in which users were logged in. This was necessary in order to obtain information on gender, which is a prerequiste for AI fairness algorithms. Seventh, we performed feature engineering. For this, we transformed the three variables (visit depth, time spent on page, and cumulative number of visits) via the following functions: average, standard deviation, 25 % quantile, maximum, and approximate entropy. We also experimented with other functions, such as 75 % quantile and minimum, but these did not result in a performance improvement and were thus discarded.

Appendix 2: Hyperparameter Tuning

Table 6 reports the tuning parameters used in our grid search. For the neural network, we used a single hidden layer. Increasing the number of layers did not improve the overall performance due to the limited size of the dataset. All parameters were estimated using Adam and early stopping.

Appendix 3: Sensitivity Analysis

In addition to boosting, all experiments are repeated based on the lasso and a deep neural network. Overall, the classifiers based on boosting yield the best prediction performance, registering an AUROC of 0.727 for the default classifier. In contrast, the default classifiers based on the lasso and the neural network correspond to 0.596 and 0.664, respectively. A possible explanation for the lower performance of the lasso is due to non-linearities in the data, which the lasso is unable to capture. Furthermore, the lower performance of the neural network is presumably due to the relatively small quantity of training data, in which case deep neural networks are unable to realize their full potential (Kraus et al. 2020). However, we consider AUROC to be particularly important, as practitioners commonly rely on the algorithm yielding the highest prediction performance. Moreover, AUROC measures the prediction performance across all possible classification thresholds and, hence, is independent of the financial costs that are specific to our setting. Therefore, the results for AUROC support our choice of boosting as the underlying prediction algorithm.

Nonetheless, the findings yielded by boosting are mainly supported by those obtained from the lasso and the neural network (Tables 3 and 4). In particular, the fair classifier results in higher financial costs across all configurations, that is, for all prediction algorithms and both notions of fairness. Similarly, in terms of fairness, the fair classifier provides a higher level than the default classifier for all configurations and both notions under study. However, only in the case of boosting and the lasso does the fair classifier provide a higher level of statistical parity than the blinded classifier. For the neural network, the fair and blinded classifiers yield a similar level. This is partly relativized by the high standard deviations that are observed for results from the neural network.

Furthermore, the fair classifier has been implemented with two additional algorithms for AI fairness, namely a preprocessing algorithm (i. e., reweighing; Kamiran and Calders 2012) and an in-processing algorithm (i. e., adversarial debiasing; Zhang et al. 2018). However, in our setting, only a post-processing algorithm (i. e., reject option based classification; Kamiran et al. 2012) has achieved fairness, i. e., statistical parity and equalized odds, respectively. Reweighing was ineffective for both training and test data, providing a level of statistical parity and equalized odds similar to that of the default classifier. Moreover, adversarial debiasing did provide a high level of

statistical parity and equalized odds on the training data, but a level similar to that of the default classifier when applied to the test data. This observation can be explained by the nature of how the different algorithms operate. If fairness is injected at an early stage of the prediction pipeline, it might be counteracted at other stages of the pipeline, especially in the context of imbalanced datasets. Hence, our findings are in line with prior research highlighting the limitations of pre- and in-processing algorithms in real-world applications (e. g., Kamiran et al. 2012).

Acknowledgements Stefan Feuerriegel acknowledges support from the Swiss National Science Foundation (Grant 197485).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abbasi A, Li J, Clifford G, Taylor H (2018) Make fairness by design part of machine learning. Harvard Bus Rev

Angwin J, Larson J, Mattu S, Kirchner L (2016) How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Barocas S, Selbst AD (2016) Big data's disparate impact. California Law Rev 104(3):671–732

Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. http://www.fairmlbook.org

Baumann A, Haupt J, Gebert F, Lessmann S (2019) The price of privacy. Bus Inf Syst Eng 61(4):413–431

Chouldechova A, Roth A (2020) A snapshot of the frontiers of fairness in machine learning. Commun ACM 63(5):82–89

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: ACM SIGKDD international conference on knowledge discovery & data mining (KDD)

Datta A, Tschantz MC, Datta A (2015) Automated experiments on ad privacy settings: a tale of opacity, choice, and discrimination. Proc Privacy Enhancing Technol 1:92–112

de Bock KW, van den Poel D (2010) Predicting website audience demographics for web advertising targeting using multi-website clickstream data. Fundamenta Informaticae 98(1):49–70

Digitec Galaxus (2018) Mobile-shopping: immer mehr Schweizer kaufen mit dem Handy im Netz ein. https://www.galaxus.ch/MWS/Release/180528_Medienmitteilung/Medienmitteilung.pdf

Ding AW, Li S, Chatterjee P (2015) Learning user real-time intent for optimal dynamic web page transformation. Inf Syst Res 26(2):339–359

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Innovations in theoretical computer science conference

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: ACM SIGKDD international conference on knowledge discovery and data mining

Feuerriegel S, Dolata M, Schwabe G (2020) Fair AI. Bus Inf Syst Eng

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2018) A comparative study of fairness-enhancing interventions in machine learning. In: Conference on fairness, accountability, and transparency (FAT)

Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci 115:E3635–E3644

GDPR (2016) Regulation EU 2016/679 of the european parliament and of the council of 27 April 2016, article 22. Off J Eur Union L 119:59

Gofman A, Moskowitz H, Mets T (2009) Integrating science into web design: consumer-driven web site optimization. J Consum Market 26:286–298

Haas C (2019) The price of fairness: a framework to explore trade-offs in algorithmic fairness

Hacker P (2018) Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law. Common Market Law Rev 55(4):1143–1185

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: International conference on neural information processing systems (NIPS), USA

Hastie T, Tibshirani R, Friedman JH (2017) The elements of statistical learning: data mining, inference, and prediction, second edition, corrected at 12th printing, 2017th edn. Springer series in statistics, Springer, New York, NY

Hatt T, Feuerriegel S (2020) Early detection of user exits from clickstream data: A Markov modulated marked point process model. In: The web conference (www)

Holstein K, Wortman Vaughan J, Daumé H, Dudik M, Wallach H (2019) Improving fairness in machine learning systems. In: Chi conference on human factors in computing systems

Jenkins P (2019) Clickgraph: Web page embedding using clickstream data for multitask learning. In: The web conference (www)

Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33(1):1–33

Kamiran F, Karim A, Zhang X (2012) Decision theory for discrimination-aware classification. In: IEEE international conference on data mining

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv:160905807

Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores

Koehn D, Lessmann S, Schaal M (2020) Predicting online shopping behaviour from clickstream data using deep learning. Expert Syst Appl 150(113):342

Kraus M, Feuerriegel S, Oztekin A (2020) Deep learning in business analytics and operations research: models, applications and managerial implications. Eur J Oper Res 281(3):628–641

Lambrecht A, Tucker C (2019) Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. Manag Sci 65(7):2966–2981

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Conference on neural information processing systems (NIPS)

Maedche A, Legner C, Benlian A, Berger B, Gimpel H, Hess T, Hinz O, Morana S, Söllner M (2019) AI-based digital assistants. Bus Inf Syst Eng 61(4):535–544

McDowell WC, Wilson RC, Kile CO (2016) An examination of retail website design and conversion rate. J Bus Res 69(11):4837–4842

Mobasher B (2007) Data mining for web personalization. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The adaptive web: Lecture notes in computer science. Springer, Heidelberg, pp 90–135

Montgomery AL, Li S, Srinivasan K, Liechty JC (2004) Modeling online browsing and path analysis using clickstream data. Market Sci 23(4):579–595

Müller O, Fay M, vom Brocke J (2018) The effect of big data and analytics on firm performance: an econometric analysis considering industry characteristics. J Manag Inf Syst 35(2):488–509

Olbrich R, Holsing C (2011) Modeling consumer purchasing behavior in social shopping communities with clickstream data. Int J Electron Commer 16(2):15–40

Reimers I, Xie C (2019) Do coupons expand or cannibalize revenue? Evidence from an e-market. Manag Sci 65(1):286–300

Russell S, Dewey D, Tegmark M (2015) Research priorities for robust and beneficial artificial intelligence. AI Mag 36(4):105–114

Senoner J, Netland T, Feuerriegel S (2021) Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. Manag Sci, forthcoming

Sheil H, Rana O, Reilly R (2018) Understanding ecommerce clickstreams: a tale of two states. In: KDD deep learning workshop

Smith M, Neupane S (2018) Artificial intelligence and human development: toward a research agenda. International Development Research Center

Statista (2020) Global online shopper conversion rate 2017-2018. https://www.statista.com/statistics/439576/online-shopper-conversion-rate-worldwide/

Susser D, Roessler B, Nissenbaum HF (2019) Online manipulation: hidden influences in a digital world. Georgetown Law Technol Rev 4(1):1–45

Valentino-Devries J, Singer-Vine J, Soltani A (2012) Websites vary prices, deals based on users' information. Wall Street J 10:60–68

Veale M, Binns R (2017) Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data Soc 4(2):1–17

White & Case (2017) Algorithms and bias: What lenders need to know. White & Case (https://www.whitecase.com/publications/insight/algorithms-and-bias-what-lenders-need-know)

Wick M, Panda S, Tristan JB (2019) Unlocking fairness: a trade-off revisited. In: Advances in neural information processing systems

Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: AAAI/ACM conference on AI, ethics, and society (AIES)