

Disparate Impact Diminishes Consumer Trust Even for Advantaged Users

Tim Draws^{1,2*} (✉), Zoltán Szlavik^{1,3*}, Benjamin Timmermans^{1,4*}, Nava Tintarev⁵, Kush R. Varshney⁴, and Michael Hind⁴

¹ IBM Center for Advanced Studies Benelux

² Delft University of Technology

t.a.draws@tudelft.nl

³ myTomorrows

zoltan.szlavik@mytomorrows.com

⁴ IBM Research

b.timmermans@nl.ibm.com, krvarshn@us.ibm.com, hindm@us.ibm.com

⁵ Maastricht University

n.tintarev@maastrichtuniversity.nl

Abstract. Systems aiming to aid consumers in their decision-making (e.g., by implementing persuasive techniques) are more likely to be effective when consumers trust them. However, recent research has demonstrated that the machine learning algorithms that often underlie such technology can act unfairly towards specific groups (e.g., by making more favorable predictions for men than for women). An undesired disparate impact resulting from this kind of algorithmic unfairness could diminish consumer trust and thereby undermine the purpose of the system. We studied this effect by conducting a between-subjects user study investigating how (gender-related) disparate impact affected consumer trust in an app designed to improve consumers’ financial decision-making. Our results show that disparate impact decreased consumers’ trust in the system and made them less likely to use it. Moreover, we find that trust was affected to the same degree across consumer groups (i.e., advantaged and disadvantaged users) despite both of these consumer groups recognizing their respective levels of personal benefit. Our findings highlight the importance of fairness in consumer-oriented artificial intelligence systems.

Keywords: disparate impact · algorithmic fairness · consumer trust

1 Introduction

Applications that seek to advise or nudge consumers into better decision-making (e.g., concerning personal health or finance) can only be effective when consumers trust their guidance. Trustworthiness is an essential aspect in the design of such *persuasive technology* (PT); i.e., technology aiming to change attitudes or behaviors without using coercion or deception [26,28,41]; because consumers

* current affiliation

are unlikely to use (or be persuaded by) systems that they do not trust [23,35]. Recent research has identified several factors that affect consumer trust in this context; including consumers' emotional states [1] as well as the system's reliability [26] and transparency [35]. Moreover, it has been argued that trust also depends on *moral expectations* that consumers have towards the technology they use [26,35]. Consumer trust could increasingly depend on such moral expectations as more systems implement machine learning algorithms (e.g., in personal health [33,35] or finance [20] applications) that make them harder to scrutinize.

A specific moral expectation that acts as a requirement for trust in this context may be *fairness* [39]. When nudges and advice are tailored to the individual consumer using machine learning, consumers may expect that the system acts fairly towards different consumer groups (e.g., concerning race or gender). Nudging or advising such that the degree of positive impact that the system has on consumers' lives *varies* with group membership could constitute an undesired *disparate impact*. For example, a *robo-advisor* (i.e., PT designed to improve consumers' financial situation [20]) could have a disparate impact by systematically recommending "safer", lower-risk investments to female consumers compared to male consumers, yielding them lower returns. Such disparate impact would violate the moral expectation of fairness and thereby undermine consumer trust.

Employing machine learning in consumer-oriented applications often holds the promise of increasing their usefulness to the individual consumer [33,46] but also bears a greater vulnerability for disparate impact. Recent research has demonstrated that machine learning algorithms may unfairly discriminate based on group membership [2,6,27]. Such discrimination is referred to as *algorithmic unfairness* if a pre-defined notion of fairness is violated [27,42] and can easily lead to an undesired disparate impact [6,13]. For example, outcomes in advice from robo-advisors may differ between groups, given that financial advice has historically been gender-biased to the disadvantage of female consumers [5,24] and algorithmic unfairness often results from disparities in the historical data that is used to train the algorithm [27]. Although several methods have been developed to mitigate algorithmic unfairness [7], in many cases it is currently not possible to do so to a satisfactory degree [8].

Disparate impact is thus a realistic issue that could undermine the efficacy of consumer-oriented artificial intelligence (AI) systems. It has been argued that fairness plays a key role in fostering trust in AI [4,34,37,39,40]. However, to the best of our knowledge, no previous work has *studied* the influence of undesired disparate impact (i.e., as a result of algorithmic unfairness) on consumer trust. It is further unclear whether unfairly advantaged consumers are affected to the same degree as disadvantaged consumers in this context. That is, the influence of disparate impact on consumer trust may depend on *perceived personal benefit* (i.e., advantaged users trusting the system's advice despite disparate impact as long as they personally benefit) or not (i.e., advantaged users losing trust in lockstep with disadvantaged users despite a perceived personal benefit). We study the effect of disparate impact on consumer trust at the use case of gender bias in robo-advisors by investigating the following research questions:

- **RQ1.** Does an apparent disparate impact of a robo-advisor affect the degree of trust that consumers place in it?
- **RQ2.** Does disparate impact affect the trust of unfairly advantaged consumers to a different degree than that of unfairly disadvantaged consumers?

To answer these questions, we conducted a between-subjects user study where we exposed participants to varying degrees of disparate impact of a robo-advisor (i.e., advantaging male users; see Section 3). Our results show that disparate impact negatively affected consumer’s trust in the robo-advisor and decreased their willingness to use it (see Section 4). Furthermore, we find that, despite both groups recognizing their respective personal (dis)advantage, *both* the disadvantaged group (women) as well as the advantaged group (men) experienced the same decrease in trust when learned about a disparate impact of the robo-advisor. Our findings underline the importance of ensuring algorithmic fairness in consumer-oriented (AI) systems when aiming to maintain consumer trust.

2 Background and Related Work

We study the effect of disparate impact on consumer trust at the use case of gender bias in robo-advisors. Our reasons for choosing the financial domain here are threefold. First, algorithmic decision-making is already widespread in consumer-oriented financial applications (e.g., in robo-advisors) [20]. Second, algorithmic decision-making in such systems is highly impactful: it directly affects consumers’ financial situations and thereby their life quality. Third, (human) financial advice has traditionally been gender-biased, underestimating and disadvantaging female consumers [24,5]. Historical data on financial advice thus contain these biases. If the algorithms that underlie robo-advisors are trained using these data, robo-advisors may have according disparate impact.

Trust in AI systems. Consumers do not use systems that they do not trust [23]. That is why trust is an important aspect in the interaction between consumer-oriented AI systems (e.g., those implementing PT) and consumers [1,26,35,41]. Recent research has linked trust in such systems to the reliability [26] and transparency [35] of the system at hand as well as consumers’ emotional states [1] and moral expectations [26,35]. Such moral expectations may gain in importance as systems increasingly rely on machine learning algorithms [20,30,33,35,46]. Moreover, whereas in some cases consumers fall prey to *automation bias* (i.e., a tendency to prefer automated over human decisions) [10], in other cases, they experience what has been referred to as *algorithm aversion*: a tendency to prefer human over algorithmic advice [11,29,32]. Research has shown that algorithm aversion can be the result of witnessing how an algorithm errs [12]. Especially in cases where a machine learning algorithm acted *unfairly*, leading to an undesired disparate impact (i.e., violating consumers’ moral expectations and reflecting erroneous decision-making), consumer trust could thus be diminished.

Measuring and mitigating algorithmic unfairness. Research has demonstrated that machine learning algorithms can make biased (unfair) predictions to the

disadvantage of specific groups [2,6,27,43]. For instance, AI systems may discriminate between white and black defendants in predicting their likelihood of re-offending [2] and between male and female consumers in predicting their creditworthiness [43]. Several methods have been proposed to measure and mitigate biases in algorithmic decision-making [7,15,21,22,27,47]. Despite these efforts, the measurement and mitigation of algorithmic bias remain challenging [8,27].

Disparate impact and trust. Algorithmic fairness has been identified as a core building block of trustworthy AI systems [4,34,39,37,40], yet few studies directly investigate the relationship between algorithmic fairness (or disparate impact) and consumer trust. Participants in one study reported that learning about algorithmic unfairness induced negative feelings and that it might cause them to lose trust in a company or product [45]. Consumers have further expressed general concerns about disparate impact of AI on a societal level [3] and are more likely to judge decisions as less fair and trustworthy if they are made by an algorithm as opposed to a human [19]. However, it has also been shown that the degree to which people are concerned about disparate impact depends on their personal biases [31,36]. What remains unclear is to what extent disparate impact (as a result of algorithmic unfairness) affects consumer trust and, if so, who (i.e., unfairly advantaged and disadvantaged consumers) are affected in particular.

3 Method

To investigate the two research questions identified in Section 1, we conducted a between-subjects user study. The setting of this study was a fictional scenario in which a bank offers a robo-advisor – called the *AI Advisor* – to its customers. We aimed to perform a granular analysis of the effect of disparate impact on consumer trust by exposing participants to different degrees of disparate impact supposedly caused by the *AI Advisor* and measuring their attitudes towards this system. Specifically, we analyzed whether the different degrees of disparate impact affected participant’s *trust* (i.e., whether they believed that the AI Advisor would make correct predictions and therefore benefit its users). To differentiate between this general notion of trust and related attitudes, we also measured *willingness to use* and *perceived personal benefit* concerning the *AI Advisor*.

3.1 Operationalization

Dependent Variables. Our experiment involved measuring participants’ attitudes towards the *AI advisor*; specifically *trust*, *willingness to use*, *perceived personal benefit*. Each variable was measured twice: once after participants saw general user statistics (Step 1; see Section 3.3) and once after participants saw gender-specific user statistics on the *AI advisor* (Step 2). We computed difference scores from these two measurements that reflected how seeing the gender-specific statistics affected participant’s attitudes as compared to their baseline attitudes.

- *Change in Trust (Continuous).* Participants rated their trust by responding to the item “In general, the AI advisor can be trusted to make correct

recommendations” on a 7-point Likert scale. We coded all responses on an ordinal scale ranging from -3 (strongly disagree) to 3 (strongly agree) and subtracted the second measurement from the first to compute the *change in trust*. Values could thus range from -6 to 6 .

- *Change in Willingness to Use (Categorical)*. Participants could respond to the item “I would personally use the AI Advisor” with either “yes” or “no”. We recorded whether their answer had changed (i.e., “yes” to “no” or vice versa) or stayed the same in the second measurement. This variable thus encompassed three categories.
- *Change in Perceived Personal Benefit (Continuous)*. Participants rated their perceived personal benefit by responding to the item “I would personally benefit from using the AI advisor” on a 7-point Likert scale. To compute the *change in perceived personal benefit*, we again subtracted the second measurement from the first. Values could thus range from -6 to 6 .

Independent Variable. Our experiment varied depending on the condition that a participant was placed in (see Section 3.3):

- *Condition*. During the experiment, we showed participants a table with user statistics of bank customers that use the *AI advisor*. These statistics, supposedly showing the average change in bank account balance for users and non-users of the *AI Advisor*, split by gender, differed depending on the condition a participant had been placed in. Each participant saw only one of four conditions: the *control* condition (in which the statistics were balanced across genders, reflecting an absence of disparate impact) or one of three experimental conditions – which we call *little bias*, *strong bias*, and *extreme bias* – that reflected varying degrees of disparate impact in favor of male consumers. Specifically, these different degrees of disparate impact represented scenarios in which female users of the *AI advisor* were disadvantaged but still benefited from using the *AI advisor* (little bias), did not benefit from the *AI advisor* (strong bias), or would in fact benefit from *not* using the *AI advisor* (extreme bias). Table 1 shows the numbers that were shown in the second statistics table in each of the conditions.

Individual Differences and Descriptive Statistics. We took two additional measurements to enable more fine-grained analyses and describe our sample:

- *Gender*. Participants could state which gender they identified with by picking from the options “male”, “female”, and “other / not specified”.
- *Age*. Participants could write their age in an open text field.

3.2 Hypotheses

Based on the research questions **RQ1** and **RQ2** introduced in Section 1, the related work from Section 2, and the experimental setup described in this section, we formulated several hypotheses. We expected that disparate impact will

Table 1. Fictional gender-specific statistics shown to participants during the second step of the study across. Only the top left two cells (concerning users of the *AI Advisor*) differed across conditions, reflecting varying degrees of disparate impact.

	Using AI Advisor	Not Using AI Advisor		Using AI Advisor	Not Using AI Advisor
Male	20%	10%	Male	25%	10%
Female	20%	10%	Female	15%	10%
All	20%	10%	All	20%	10%
<i>Control condition</i>			<i>Little bias condition</i>		
	Using AI Advisor	Not Using AI Advisor		Using AI Advisor	Not Using AI Advisor
Male	30%	10%	Male	35%	10%
Female	10%	10%	Female	5%	10%
All	20%	10%	All	20%	10%
<i>Strong bias condition</i>			<i>Extreme bias condition</i>		

decrease consumer trust (H1a) and that consumers will be less likely to use the *AI Advisor* (H1b) if it has disparate impact (i.e., the stronger the disparate impact, the lower consumer trust and willingness to use the *AI Advisor*). We predicted that disparate impact would affect the perceived personal benefit of male consumers differently compared to female consumers (i.e., following what the displayed statistics suggest; H2a). Accordingly, we further expected that the decrease in trust described in H1a would be moderated by gender (H2b). That is, we predicted that the trust of advantaged consumers (i.e., men) would be affected differently compared to disadvantaged consumers (i.e., women).

- **H1a.** Consumers who are exposed to statistics that reveal a disparate impact of a robo-advisor in favor of male users will *trust this system less* to give correct recommendations compared to consumers who are exposed to balanced statistics.
- **H1b.** Consumers who are exposed to statistics that reveal a disparate impact of a robo-advisor in favor of male users will be *less likely to use this system* compared to consumers who are exposed to balanced statistics.
- **H2a.** The effect of statistics suggesting a disparate impact of a robo-advisor in favor of men on *perceived personal benefit* is moderated by gender.
- **H2b.** The effect of statistics suggesting a disparate impact of a robo-advisor in favor of men on *consumer trust* is moderated by gender.

3.3 Procedure

We set up our user study by creating a task on the online study platform *Figure Eight*.⁶ Before commencing with the experiment, participants were shown a short

⁶ Since conducting this study in June 2019, *Figure Eight* has been renamed to *Appen*. More information can be found at <https://appen.com>.

introduction and asked to state their gender and age. The experiment consisted of two steps. Whereas Step 1 was the same for all participants, Step 2 differed depending on which one of four conditions a participant had been assigned to.

Step 1. We introduced participants to a fictional scenario in which they could activate a robo-advisor – called the *AI advisor* – in their banking app:

“Imagine your bank offers a digital assistant called the ‘AI advisor’. If you activate the AI advisor in your banking app, it will monitor your financial situation and give you relevant recommendations that may improve your financial situation. For example, it may suggest saving strategies or recommend investments.”

Additionally, to promote the idea that the AI Advisor is generally reliable, participants were given an idea of whether people benefit from using the *AI Advisor*:

“Overall statistics suggest that people benefit from using the AI advisor. The bank account balance of bank customers who use the AI advisor increases by an average of 20% every year, whereas the balance of customers who don’t use the AI advisor increases by an average of only 10% per year.”

Below was a table displaying the mentioned statistics. We then measured trust, willingness to use, and perceived personal benefit concerning the *AI Advisor*.

Step 2. Participants were led to a new page for the second step of the experiment. Here we added some additional information on the AI advisor:

“Next to general statistics on all bank customers, we can also look at how the AI advisor performs for subgroups of bank customers. Below you can see the change in bank account balance for men and women in particular.”

Below this text was a table similar to the table in Step 1, but with two added rows that showed the average change in bank account balance per year for men and women in particular (see Table 1). Whereas the statistics for all bank customers overall, as well as for men and women *not* using the *AI advisor* was the same in all conditions, the statistics for men and women *using* the *AI advisor* varied depending on the condition they were assigned to (see Section 3.1). Each Table 1 shows the displayed statistics for male and female users per condition. We then again measured trust, willingness to use, and perceived personal benefit.

3.4 Statistical Analyses

Testing H1a and H2b. To test whether there is an effect of disparate impact on consumer trust (H1a) that is moderated by gender (H2b), we conducted a classical ANOVA with *condition* and *gender* as between-subjects factors and *change in trust* as the dependent variable. A significant main effect of *condition* on *change in trust* in this analysis would suggest that *change in trust* differed between conditions (H1a). In this case, we would perform posthoc analyses to investigate the differences between the conditions in more detail. A significant interaction effect between *condition* and *gender* would suggest that the conditions had a different effect for the disadvantaged group (i.e. female participants) compared to the advantaged group (i.e., male participants; H2b).

We further conducted a Bayesian ANOVA according to the protocol proposed by van den Bergh et al. [38]. Bayesian hypothesis tests involve the computation of the *Bayes factor*, a quantitative comparison of the predictive power of two competing statistical models [44]. The Bayes factor weighs the evidence provided by the data and thus allows for direct model comparison. Practically, comparing different models (i.e., including or excluding an interaction effect of *condition* and *gender*) this way allowed for a richer interpretation of our results. We performed the Bayesian ANOVA using the software JASP [16] with default settings. We computed *Bayes Factors* (BFs) by comparing the models of interest to a *null model*⁷ and interpreted them according to the guidelines proposed by Lee and Wagenmakers [18], who adopted them from Jeffreys [17].

Testing H1b. We tested whether disparate impact affected participants’ willingness to use the *AI Advisor* by conducting a chi-squared test between *condition* and *change in willingness to use*. A significant result in this analysis would suggest that the number of participants’ who changed their willingness to use the *AI Advisor* differed across conditions.

Testing H2a. We conducted another ANOVA with *condition* and *gender* as between-subjects factors and *change in perceived personal benefit* as dependent variable to test whether *gender* acted as a moderator here. A significant interaction effect in this analysis would indicate that this was the case.

Significance Threshold and Correction for Multiple Testing. In all classical analyses we conducted, we aimed for a type 1 error probability of no more than 0.05. However, by conducting our planned analyses we automatically tested a total of seven hypotheses: three in each ANOVA (i.e., two main effects and one interaction) and one in the chi-squared test. This meant that the probability of committing a type 1 error rose considerably [9]. Therefore, we adjusted our significance threshold by applying a Bonferroni correction, where the desired type 1 error rate is divided by the number of hypotheses that are tested [25]. In our main analyses we thus handled a significance threshold of $\frac{0.05}{7} = 0.007$ and only regarded results as statistically significant if their *p*-value fell below this adjusted threshold. The same procedure was applied for posthoc analyses comparing each of the four conditions with each other as this meant conducting $\binom{4}{2} = 6$ hypothesis tests (i.e., adjusting the threshold to $\frac{0.05}{6} = 0.008$).

3.5 Participants

We recruited 567 participants via the *Figure Eight* pool of contributors (554) and direct contacts (13). Seventy-three participants were excluded from the study because they either filled at least one of the obligatory text fields with less than 10 characters, took less than 60 seconds to complete the task, or took more than 10 minutes to complete the task. Furthermore, we did not analyze data of five participants who stated “other / not specified” as their gender because our study involved a disparate impact between male and female consumers.

⁷ The *null model* in this procedure consisted of only an intercept.

After exclusion, 489 participants remained. Of those, 238 (49%) were male and 251 (51%) were female; with a mean age of 41.9 (sd = 13.1). Participants recruited by Figure Eight received \$0.10 as payment for participation. Random allocation to the four conditions resulted in 124, 121, 121, and 122 participants in the *control*, *little bias*, *strong bias*, and *extreme bias* conditions, respectively.

4 Results

H1a: Disparate Impact Decreased Consumer Trust. As hypothesized, *change in trust* differed across conditions ($F = 6.906$, $p < 0.001$; see the left-hand panel of Figure 1). The results from the Bayesian ANOVA confirm this result, showing strong evidence for a main effect of condition ($BF_{10} = 70.02$, see Table 2). To test for differences between the individual conditions, we conducted posthoc analyses (i.e., Mann-Whitney U tests). Only the difference between the *control* and *extreme bias* conditions was significant ($W = 9368$, $p < 0.001$). This suggests that participants lost trust due to disparate impact, but also that the unfairness needed to be comparatively extreme for this effect to occur.

Table 2. Bayesian ANOVA with *change in trust* as dependent variable.

Models	P(M)	P(M Data)	BF_M	BF_{10}	Error %
Null model	0.200	8.209e-4	0.003	1.000	
condition	0.200	0.057	0.244	70.024	0.001
gender	0.200	0.004	0.018	5.413	1.533e-6
condition + gender	0.200	0.735	11.116	895.772	1.638
condition + gender + condition * gender	0.200	0.202	1.012	245.894	1.978

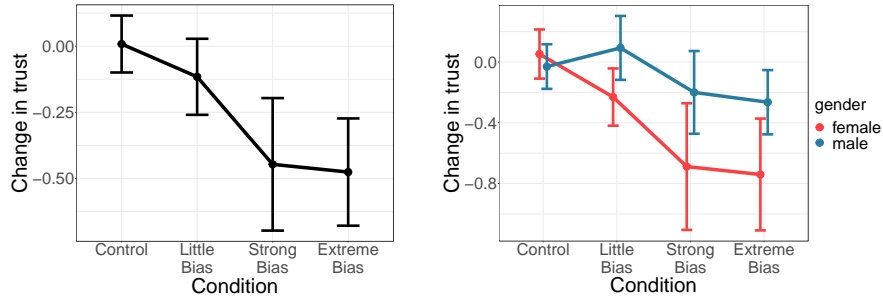


Fig. 1. *Change in trust* across conditions for all participants (left-hand panel) and split by gender (right-hand panel). The error bars represent 95% confidence intervals.

H1b: Disparate Impact Decreased Willingness to Use. In accordance with disparate impact negatively affecting trust (H1a), it decreased participants' willingness to use the *AI Advisor* (see Table 3). The increasing proportion of participants who changed their attitude from “yes” to “no” as conditions reflected stronger disparate impact was statistically significant ($\chi^2 = 25.06$, $p < 0.001$).

Table 3. *Change in willingness to use across conditions.* The labels $-$, $=$, and $+$ reflect changes from “yes” to “no”, no change, “no” to “yes”, respectively.

Change	Condition			
	Control	Little Bias	Strong Bias	Extreme Bias
$-$	1	9	23	16
$=$	121	111	97	105
$+$	2	1	1	1
total	124	121	121	122

H2a: Gender Moderated the Effect of Disparate Impact on Perceived Personal Benefit. As expected, the results from the second ANOVA show a significant interaction effect of *condition* and *gender* on *change in perceived personal benefit* ($F = 8.525$, $p < 0.001$). This means that male participants’ perceived personal benefit was affected differently compared to that of female participants. More specifically, Figure 2 shows that whereas men’s perceived personal benefit did not change due to seeing the gender-specific user statistics across conditions, female participants perceived increasingly lower levels of personal benefit as disparate impact (to their disadvantage) became more severe.

H2b: Male Consumers Experienced the Same Decrease in Trust as Female Consumers. In contrast to what we hypothesized, we do not find a significant interaction effect of *condition* and *gender* on *change in trust* ($F = 2.094$, $p = 0.096$; see the right-hand panel of Figure 1). We can therefore not conclude that the conditions had a different effect on male participant’s *change in trust* compared to that of female participants. The Bayesian ANOVA confirms this result: the model containing just two main effects for condition and gender explain the data best ($BF_{10} = 895.77$; see Table 2) and roughly four times better than the model that includes the interaction effect ($BF_{10} = 245.89$). This suggests that unfairly advantaged and disadvantaged participants (i.e., men and women, respectively) experienced the same decrease in trust due to algorithmic unfairness despite diverging levels of perceived personal benefit (H2a).

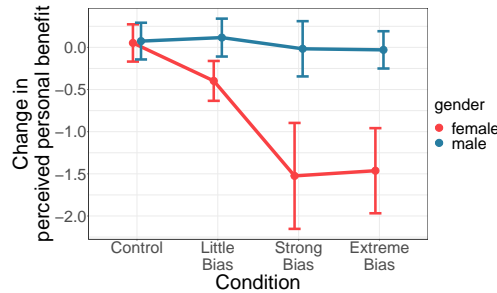


Fig. 2. *Change in perceived personal benefit across conditions and split by gender.* The error bars represent 95% confidence intervals.

5 Discussion

In this paper, we presented a between-subjects user study that aimed to investigate the influence of algorithmically-driven disparate impact on consumer trust at the use case of gender-bias in robo-advisors. Our results suggest that disparate impact – at least when it is extreme – decreases trust and makes consumers less likely to use such systems. We further find that, although disadvantaged and advantaged users recognize their respective levels of personal benefit in scenarios of disparate impact, both experience equally decreasing levels of trust when they learn about a disparate impact caused by the system at hand. Our work contributes to a growing body of literature that highlights the importance of ensuring fairness and avoiding disparate impact of consumer-oriented AI systems.

5.1 Implications

Our findings have implications for consumers as well as industry. Consumers should be aware that machine-learning-based applications can be biased. If disparate impact is an important factor for consumer trust, consumers need to think critically when using such systems. One potential way forward for consumers would be to demand from companies to publish independently carried out research into the (algorithmic) fairness and impact of their products.

Publishers of consumer-oriented AI systems need to establish algorithmic fairness in their products and avoid disparate impact to serve consumers effectively. Our findings show that failing to do so may lead to a decrease in consumers’ trust and willingness to use such systems.

5.2 Limitations and Future Work

Our study is subject to at least five important limitations. First, we studied the effect of disparate impact on consumer trust at a specific use case: a binary gender bias in robo-advisors. This makes our results difficult to generalize because many other forms of bias (including those based on race, religion, or sexual orientation) as well as other AI systems (e.g., for recommendations of medical treatment, tourist attractions, or movies) exist. It is easy to imagine how consumer trust could be affected differently when, for example, disparate impact concerns small minorities, multitudes of gender identities (or another consumer characteristic), a chosen group membership such as consumers’ profession, or a system that is less impactful on consumers’ personal lives than a robo-advisor. On a related note, we here positioned women in the disadvantage and men in the advantage (i.e., the setting that corresponds to biases in human financial advice) but it is not certain if we were to obtain the same results if the (dis-)advantage was distributed the other way round. Future work could explore these different scenarios to help generalize and better understand the relationship between the effect of disparate impact on consumer trust.

Second, our finding that advantaged and disadvantaged users experienced the same decrease in trust appears to go counter to previous research suggesting that

people make stronger fairness judgments when they are personally affected [14]. However, it is not clear from our results to what degree advantaged users (i.e., men) felt personally affected; e.g., because they have women in their lives who they deeply care about. The role of personal relevance in the effect of disparate impact on consumer trust thus remains to be clarified by future research.

Third, our results show a decreasing trend in consumer trust as conditions become more extreme, but show a statistically significant difference only between the control and extreme bias conditions. Future work could examine these differences (also across domains) in more detail to establish the relationship between *the level of* disparate impact and consumer trust (e.g., to determine what lies within and beyond an “acceptable margin” of disparate impact).

Fourth, we studied fairness related to group membership (i.e., gender), which might elicit a different (moral) evaluation than fairness on the individual level. Our results show that trust can decrease despite perceived personal benefit. However, this effect might have been caused by a sense of loyalty towards the disadvantaged group. An interesting direction for future work is to study whether similar patterns emerge when disparate impact concerns individuals; e.g., when advantaged and disadvantaged subjects are randomly chosen.

6 Conclusion

We presented a user study investigating the effect of algorithmically-driven disparate impact (i.e., when algorithm outcomes adversely affect one group of consumers compared to another) on consumer trust. Specifically, we studied the effect of *gender-bias* in an application that aimed to persuade consumers’ to make better financial decisions. We found that disparate impact decreased participants’ trust and willingness to use the application. Furthermore, our results show that the trust of unfairly advantaged participants was just as affected as that of disadvantaged participants. These findings imply that disparate impact (i.e., as a result of algorithmic unfairness) can undermine trust in consumer-oriented AI systems and should therefore be avoided or mitigated when aiming to create trustworthy technology.

Acknowledgements

This research has been supported by the *Think Forward Initiative* (a partnership between ING Bank, Deloitte, Dell Technologies, Amazon Web Services, IBM, and the Center for Economic Policy Research – CEPR). The views and opinions expressed in this paper are solely those of the authors and do not necessarily reflect the official policy or position of the Think Forward Initiative or any of its partners.

References

1. Ahmad, W.N.W., Ali, N.M.: A Study on Persuasive Technologies: The Relationship between User Emotions, Trust and Persuasion. *Int. J. Interact. Multimed. Artif. Intell.* **5**(1), 57–61 (2018). <https://doi.org/10.9781/ijimai.2018.02.010>
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There’s software used across the country to predict future criminals and it’s biased against blacks. *ProPublica* (2019), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
3. Araujo, T., Helberger, N., Kruikemeier, S., de Vreese, C.H.: In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI Soc.* **35**(3), 611–623 (2020). <https://doi.org/10.1007/s00146-019-00931-w>
4. Arnold, M., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., Varshney, K.R., Bellamy, R.K., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K.N., Olteanu, A.: FactSheets: Increasing trust in AI services through supplier’s declarations of conformity. *IBM J. Res. Dev.* **63**(4-5) (2019). <https://doi.org/10.1147/JRD.2019.2942288>
5. Baeckström, Y., Silvester, J., Pownall, R.A.: Millionaire investors: financial advisors, attribution theory and gender differences. *Eur. J. Financ.* **24**(15), 1333–1349 (2018). <https://doi.org/10.1080/1351847X.2018.1438301>
6. Barocas, Solon and Selbst, A.D.: Big data’s disparate impact. *Calif. Law Rev.* **104**(671), 671–732 (2016)
7. Bellamy, R.K., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S.: AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* **63**(4-5) (2019). <https://doi.org/10.1147/JRD.2019.2942287>
8. Corbett-Davies, S., Goel, S.: The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv Prepr. arXiv1808.00023* (2018), <http://arxiv.org/abs/1808.00023>
9. Cramer, A.O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R.P., Waldorp, L.J., Wagenmakers, E.J.: Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychon. Bull. Rev.* **23**(2), 640–647 (2016). <https://doi.org/10.3758/s13423-015-0913-5>
10. Cummings, M.L.: Automation bias in intelligent time critical decision support systems. *Collect. Tech. Pap. - AIAA 1st Intell. Syst. Tech. Conf.* **2**, 557–562 (2004). <https://doi.org/10.2514/6.2004-6313>
11. Diab, D.L., Pui, S.Y., Yankelevich, M., Highhouse, S.: Lay perceptions of selection decision aids in US and non-US samples. *Int. J. Sel. Assess.* **19**(2), 209–216 (jun 2011). <https://doi.org/10.1111/j.1468-2389.2011.00548.x>
12. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**(1), 114–126 (2015). <https://doi.org/10.1037/xge0000033>
13. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and Removing Disparate Impact. In: *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* pp. 259–268 (2015)
14. Ham, J., van den Bos, K.: Not fair for me! The influence of personal relevance on social justice inferences. *J. Exp. Soc. Psychol.* **44**(3), 699–705 (2008). <https://doi.org/10.1016/j.jesp.2007.04.009>

15. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. *Adv. Neural Inf. Process. Syst.* pp. 3323–3331 (2016)
16. JASP Team: JASP (Version 0.14) (2020)
17. Jeffreys, H.: *Theory of Probability*. Oxford University Press, Oxford (1939)
18. Lee, M.D., Wagenmakers, E.J.: *Bayesian cognitive modeling: A practical course*. Cambridge University Press (2014). <https://doi.org/10.1017/CBO9781139087759>
19. Lee, M.K.: Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data Soc.* **5**(1), 1–16 (2018). <https://doi.org/10.1177/2053951718756684>
20. Lieber, R.: Financial Advice for People Who Aren't Rich (apr 2014), <https://www.nytimes.com/2014/04/12/your-money/start-ups-offer-financial-advice-to-people-who-arent-rich.html>
21. Mary, J.J., Calauzènes, C., Karoui, N.E.: Fairness-Aware Learning for Continuous Attributes and Treatments. *Icml* **97**, 4382–4391 (2019), <http://proceedings.mlr.press/v97/mary19a.html>
22. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. *arXiv Prepr. arXiv1908.09635* (2019), <http://arxiv.org/abs/1908.09635>
23. Muir, Bonnie M and Moray, N.: Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* **39**(3), 429–460 (1996)
24. Mullainathan, S., Noeth, M., Schoar, A.: The Market for Financial Advice: An Audit Study. *SSRN Electron. J.* (2012). <https://doi.org/10.2139/ssrn.1572334>
25. Napierala, M, A.: What Is the Bonferroni correction? (2012), <http://www.aaos.org/news/aaosnow/apr12/research7.asp>
26. Nickel, P., Spahn, A.: Trust, Discourse Ethics, and Persuasive Technology. In: *Persuas. Technol. Des. Heal. Safety; 7th Int. Conf. Persuas. Technol.* 2012. pp. 37–40. Linköping University Electronic Press (2012)
27. Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdil, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., Staab, S.: Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **10**(3), 1–14 (2020). <https://doi.org/10.1002/widm.1356>
28. Oinas-Kukkonen, H., Harjumaa, M.: Towards deeper understanding of persuasion in software and information systems. In: *Proc. 1st Int. Conf. Adv. Comput. Interact. ACHI 2008* (2008). <https://doi.org/10.1109/ACHI.2008.31>
29. Önkal, D., Goodwin, P., Thomson, M., Gönül, S., Pollock, A.: The relative influence of advice from human experts and statistical methods on forecast adjustments. *J. Behav. Decis. Mak.* **22**(4), 390–409 (2009). <https://doi.org/10.1002/bdm.637>
30. Orji, R., Moffatt, K.: Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health Informatics J.* **24**(1), 66–91 (2018). <https://doi.org/10.1177/1460458216650979>
31. Otterbacher, J., Checco, A., Demartini, G., Clough, P.: Investigating user perception of gender bias in image search: The role of sexism. *41st Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 2018* pp. 933–936 (2018). <https://doi.org/10.1145/3209978.3210094>
32. Promberger, M., Baron, J.: Do patients trust computers? *J. Behav. Decis. Mak.* **19**(5), 455–468 (2006). <https://doi.org/10.1002/bdm.542>

33. Purpura, S., Schwanda, V., Williams, K., Stubler, W., Sengers, P.: Fit4Life: The Design of a Persuasive Technology Promoting Healthy Behavior and Ideal Weight. In: Proc. SIGCHI Conf. Hum. factors Comput. Syst. pp. 423–432 (2011)
34. Rossi, F.: Building trust in artificial intelligence. *J. Int. Aff.* **72**(1), 127–133 (2019)
35. Sattarov, F., Nagel, S.: Building trust in persuasive gerontechnology: User-centric and institution-centric approaches. *Gerontechnology* **18**(1), 1–14 (2019). <https://doi.org/10.4017/gt.2019.18.1.001.00>
36. Smith, J., Sonboli, N., Fiesler, C., Burke, R.: Exploring User Opinions of Fairness in Recommender Systems. In: CHI'20 Work. Human-Centered Approaches to Fair Responsible AI (2020), <http://arxiv.org/abs/2003.06461>
37. Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., van Moorsel, A.: The relationship between trust in AI and trustworthy machine learning technologies. FAT* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar. pp. 272–283 (2020). <https://doi.org/10.1145/3351095.3372834>
38. Van Den Bergh, D., Van Doorn, J., Marsman, M., Draws, T., Van Kesteren, E.J., Derks, K., Dablander, F., Gronau, Q.F., Kucharský, Š., Gupta, A.R.N., Sarafoglou, A., Voelkel, J.G., Stefan, A., Ly, A., Hinne, M., Matzke, D., Wagenmakers, E.J.: A tutorial on conducting and interpreting a bayesian ANOVA in JASP. *Annee Psychol.* **120**(1), 73–96 (2020). <https://doi.org/10.3917/anpsy1.201.0073>
39. Varshney, K.R.: Trustworthy machine learning and artificial intelligence. *XRDS Crossroads, ACM Mag. Students* **25**(3) (2019). <https://doi.org/10.1145/3313109>
40. Varshney, K.R.: On Mismatched Detection and Safe, Trustworthy Machine Learning. In: 2020 54th Annu. Conf. Inf. Sci. Syst. CISS 2020 (2020). <https://doi.org/10.1109/CISS48834.2020.1570627767>
41. Verbeek, P.P.: Persuasive Technology and Moral Responsibility Toward an ethical framework for persuasive technologies. *Persuasive* **6**, 1–15 (2006)
42. Verma, S., Rubin, J.: Fairness definitions explained. In: Proc. Int. Work. Softw. Fairness. pp. 1–7. FairWare '18, Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3194770.3194776>
43. Vigdor, N.: Apple card investigated after gender discrimination complaints. *New York Times* (2019)
44. Wagenmakers, E.J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q.F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J.N., Morey, R.D.: Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon. Bull. Rev.* **25**(1), 35–57 (2018). <https://doi.org/10.3758/s13423-017-1343-3>
45. Woodruff, A., Fox, S.E., Rousso-Schindler, S., Warsaw, J.: A qualitative exploration of perceptions of algorithmic fairness. *Conf. Hum. Factors Comput. Syst. - Proc.* **2018-April**, 1–14 (2018). <https://doi.org/10.1145/3173574.3174230>
46. Yang, Q., Banovic, N., Zimmerman, J.: Mapping machine learning advances from HCI research to reveal starting places for design innovation. *Conf. Hum. Factors Comput. Syst. - Proc.* **2018-April**, 1–11 (2018). <https://doi.org/10.1145/3173574.3173704>
47. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. 26th Int. World Wide Web Conf. WWW 2017 pp. 1171–1180 (2017). <https://doi.org/10.1145/3038912.3052660>