LEIBNIZ INSTITUTE
SAFE
Sustainable Architecture
for Finance in Europe

Kevin Bauer | Nicolas Pfeuffer | Benjamin M. Abdel-Karim | Oliver Hinz | Michael Kosfeld

The Terminator of Social Welfare? The Economic Consequences of Algorithmic Discrimination

SAFE Working Paper No. 287

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

The Terminator of Social Welfare?

The Economic Consequences of Algorithmic Discrimination

Kevin Bauer[1], Nicolas Pfeuffer,[2] Benjamin M. Abdel-Karim,[2]
Oliver Hinz,[2] Michael Kosfeld,[3]

Abstract

Using experimental data from a comprehensive field study, we explore the causal effects of algorithmic discrimination on economic efficiency and social welfare. We harness economic, game-theoretic, and state-of-the-art machine learning concepts allowing us to overcome the central challenge of missing counterfactuals, which generally impedes assessing economic downstream consequences of algorithmic discrimination. This way, we are able to precisely quantify downstream efficiency and welfare ramifications, which provides us a unique opportunity to assess whether the introduction of an AI system is actually desirable. Our results highlight that AI systems' capabilities in enhancing welfare critically depends on the degree of inherent algorithmic biases. While an unbiased system in our setting outperforms humans and creates substantial welfare gains, the positive impact steadily decreases and ultimately reverses the more biased an AI system becomes. We show that this relation is particularly concerning in selective-labels environments, i.e., settings where outcomes are only observed if decision-makers take a particular action so that the data is selectively labeled, because commonly used technical performance metrics like the precision measure are prone to be deceptive. Finally, our results depict that continued learning, by creating feedback loops, can remedy algorithmic discrimination and associated negative effects over time.

[1] Leibniz Institute for Financial Research SAFE, Theodor-W.-Adorno-Platz 3, D-60323 Frankfurt am Main, Germany, E-Mail: Bauer@safe.uni-frankfurt.de.

[2] Chair of Information Systems and Information Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Germany, E-Mail: Pfeuffer@wiwi.uni-frankfurt.de, Abdel-Karim@wiwi.uni-frankfurt.de, Hinz@wiwi.uni-frankfurt.de.

[3] Chair of Organization and Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Germany, E-Mail: Kosfeld@econ.uni-frankfurt.de.

°We gratefully acknowledge research support from the Leibniz Institute for Financial Research SAFE.

Introduction

The field of Artificial Intelligence (AI), especially in the area of machine learning (ML), has seen dramatic progress in the last decade (LeCun et al., 2015). Today, the use of AI systems to augment human decision-making, or even replace the human decision-maker at all, has become an integral part of daily work. At its core, the majority of current systems comprises ML algorithms that revolve around learning representations. This is done by deriving flexible mathematical functions from training data that comprises examples of input-output pairs. In that sense, ML methods can be interpreted as a very powerful tool for data-driven model selection (Domingos, 2012). Thereby models are intended to generate accurate predictions about a variable of interest (label) using available data (features) not included in the training data (Mullainathan & Spiess, 2017). Generated predictions can then be used to inform decision-making under uncertainty and environments of asymmetric information (Agrawal et al., 2019).

Against the background that their predictions are faster, cheaper, (most of the time) more reliable and scalable than human ones, AI technologies have found their way into businesses in virtually all areas of industry (McAfee et al., 2012). In the financial sector, where credit card fraud is a profound problem creating substantial economic harm (Nilson, 2016), credit card providers use ML models to predict the legitimacy of a transaction using its characteristics and data of previous transactions. Based on the prediction, an information system subsequently permits or rejects the transaction (see for example Bhattacharyya et al., 2011; Adewumi & Akinyelu, 2017).

Relatedly, there is increasing use of ML algorithms in the banking sector, where AI systems enable the accurate detection and management of risks (Leo et al., 2019). On an individual level, for instance, ML algorithms make use of historic customer data to predict applicants' risk of credit default, classify them as good or bad, and ultimately decide about granting a credit (Wang et al., 2015).

AI applications also frequently augment or automate hiring and promotion decisions in organizations by identifying individuals who are most capable of filling specific vacancies

(Hoffman et al., 2018). In this context, algorithms use available data, such as people's personal information, to produce predictions about their future performance and job fit, for both, new applicants or current employees. By informing central HR decisions with accurate individual-level predictions, AI systems promise increases in organizations' labor productivity as candidates are more likely to be matched with suitable jobs.

Other examples of AI systems augmenting or automating human decision making include algorithmic trading (Hendershott et al., 2011; Chaboud et al., 2014), predictive policing (Ensign et al., 2017), bail decisions (Kleinberg, Lakkaraju, et al., 2018), medical diagnosis (Esteva et al., 2019), and even online dating (Hitsch et al., 2010). Taken together these examples illustrate the broad adoption of and reliance on algorithmic decision making in business practice.

While all these instances foreshadow that AI systems may substantially enhance economic efficiency and social welfare, there is also the risk that algorithmic decision making may unintentionally and unexpectedly shape societal outcomes for the worse (for a comprehensive discussion see Rahwan et al., 2019). There already exists ample empirical evidence showing how the broad use of algorithms can inefficiently impose less favorable treatment to already disadvantaged groups creating societal tensions and welfare losses (Sweeney, 2013; Ensign et al., 2017; Obermeyer et al., 2019; Lambrecht & Tucker, 2019). When deciding upon the deployment of AI systems to augment or automate human decisions, we need to consider the entire range of complex consequences, both positive and negative ones and balance them. It is therefore crucial to further our understanding of how the use of AI systems scales into society-wide consequences.

With the paper at hand, we intend to contribute to this necessity. Specifically, using a controlled experimental setting we test whether letting an AI system make decisions in a strategic setting under asymmetric information leads to better individual and social outcomes compared to a human benchmark. We are mainly interested in identifying how inherent algorithmic biases shape these outcomes. Therefore, we vary the degree of bias the AI system exhibits against women and measure corresponding efficiency and welfare changes. The central challenge when it comes to evaluating economic ramifications of using AI systems lies mainly in assessing whether the AI system's decisions are better than the alternative, i.e., whether they outperform those taken by humans. One of the main problems is that one virtually never observes the consequences of an alternative decision that had not been made, i.e., there is a lack of counterfactual observations. As a consequence, it is almost impossible to assess the welfare ramifications of letting an AI system decide instead of a human when they opt for different choices. For instance, if a human decision-maker chose not to hire an applicant while a corresponding AI system would have done so, it is not possible to measure whether the algorithmic decision would have been better simply because there is no data on the applicant's performance had he been hired.

In our study, we overcome the problem of missing counterfactuals by making use of experimental data that we collected in a controlled and incentivized field study. Participants in our study answered a broad set of survey items on demographics, socio-economic background, cognitive abilities, and personality traits. Most important, participants also engaged in an incentivized sequential prisoners' dilemma, an experimental and game-theoretic paradigm mimicking the fundamental structure of many real-life situations where people make strategic decisions under uncertainty due to asymmetric information. Examples include employer-employee relations (Akerlof, 1982), principal-agent exchanges (Fehr et al., 1997), or market transactions (Fehr et al., 1993; Brown et al., 2004). The basic structure of the experimental game is as follows. There are two players - a trustor and a trustee. Both are initially endowed with 10 monetary units. First, the trustor decides whether or not to transfer his endowment to the trustee. The trustee learns about the trustor's choice and subsequently decides about a transfer of her initial endowment as well. In case of a transfer, the monetary units sent from one player to the other are doubled. This abstract setting mirrors the essence of any sequential economic exchange that takes place in the absence of perfect enforcement mechanisms. We elicited subjects' prisoners' dilemma choices using the strategy method, i.e., in the role of the trustee participants make conditional decisions for both possible decisions of the trustor. Hence, the strategy method gives us the unique opportunity to observe consequences of counterfactual choices that trustors did not make.

Using the data from our field study, we build an AI system that makes initial trustor decisions on behalf of human stakeholders who, instead of playing as the trustor themselves, delegate the decision authority to the machine. The AI system comprises two central components. First, a ML algorithm trained to predict a trustee's likelihood of reciprocating a transfer of endowment by transferring the personal endowment as well. Second, an algorithm that uses the prediction in combination with the human stakeholder's estimated preferences to make the utility-maximizing trustor decision. With the AI system, we study whether a population of subjects is better off in case an AI instead of a human makes strategic trustor choices. Given that we observe counterfactual trustee decisions, we are able to precisely measure performance differences between the AI system and human trustors in terms of individual and population-wide economic efficiency and welfare. We first study how an unbiased AI system performs relative to the human benchmark. Subsequently, we study how these results change in response to introducing different degrees of an algorithmic bias against women. We induce biases by using non-representative training data, a problem very relevant in practice. Finally, inspired by notions from papers that study ML in non-stationary environments (Elwell & Polikar, 2011), we examine whether continued learning - the ongoing updating of ML models using newly collected training examples - can help counteracting originally learned biases over time.

There are three main insights from our study. First, we provide causal evidence that AI systems' capabilities to improve economic efficiency and social welfare (on both an individual and a population-wide level) critically depends on the absence of inherent algorithmic biases against specific subgroups. The more biased an AI system is, the more it fosters the occurrence of inefficient outcomes and reduces welfare on both, the individual and the social level. The size of negative ramifications increases with inherent biases. Notably, even the group against which the AI system does not discriminate is better off if the predictive ML component did not inherit a bias from non-representative training data. Second, we depict that in settings prone to selective labels issues (Lakkaraju et al., 2017) the observed algorithmically shaped outcomes only allow to construct poor technical performance measures for the employment of the machines. Independent of their inherent biases and welfare consequences, the selectively observed outcomes suggest that all AI systems perform equally well with respect to technical performance metrics. This is the case even though strongly biased systems create considerable welfare losses which we can only observe in our study because we have access to counterfactuals that are usually not accessible in business practice and most of real life settings. These insights suggest that algorithmically created welfare losses in selective labels environments may remain undetected for a long time and emphasizes the importance of consulting non-technical performance measures when assessing the efficacy of AI systems. Finally, we demonstrate that continued learning in a stable environment where there is no discrimination can, at least to some extent, repair originally biased algorithms. The introduction of an updating apparatus creates feedback effects through which initially distortions in the training data increasingly vanish. Retraining the ML algorithm on more and more representative training data increases its predictive performance considerably over time. This findings indicates that there can be a benefit to ensuring the continued maintenance and controlled updating of AI systems in practice.

The paper proceeds as follows. In section 2, we summarize related literature. Section 3 develops a game-theoretic framework that serves as a formal illustration of how the use of an AI system may shape population-wide outcomes in terms of efficiency and welfare. We explain details of the conducted field study, the structure of the data, and the simulation exercises in section 4. Section 5 presents our results. Finally, section 6 discusses findings and concludes.

Related Literature

Our study aims to document causal efficiency and welfare consequences from letting differently biased AI systems instead of humans make strategic decisions under uncertainty. To this end we choose an intentionally abstract sequential exchange setting enabling us to observe ramifications of counterfactual choices. This is, consequences of choices that have not actually been made. We measure economic ramifications of introducing biased systems on both individual and social levels. In our setting, AI systems possess different degrees of a bias against women. We use gender as an example for a broad class of characteristics that algorithms can base discrimination on (e.g. ethnic background, religion, sexual orientation), but we have no access to in our data. With this objective, the article at hand contributes to three distinct streams of literature.

The first and most closely related line of work is a nascent literature concerned with the consequences of employing AI systems to augment or automate human decision-making. In the context of medical diagnosing, Mullainathan and Obermeyer (2017) argue that the use of predictive ML algorithms as a decision aid can amplify existing moral hazard and policy problems in the health system, in case they are naively trained on data prone to measurement errors. Therefore, the efficacy of employing algorithmic decision support systems depends case-by-case on the design and structure of algorithms and may not generally augment social welfare. In a forward-looking assessment of the potential impact of AI systems on economics, Athey (2018) argues that ML-powered technologies not only possess the potential to create immediate efficiency gains but that their use may also entail more complex downstream ramifications. Illustrating the complexity in assessing the total welfare consequences, Athey conjectures that considerable decreases in transportation costs caused by the use of autonomous vehicles may also decrease the housing costs for people who live in commuting distance of cities. Kleinberg, Lakkaraju, et al. (2018) studies whether an algorithmic decision aid can improve judges' bail decisions by providing a prediction about a defendant's recidivism risk. Using a data set on pretrial bail decisions of different judges and econometric proxies to circumvent the missing counterfactuals problem, the authors produce evidence that machine learning applications can lead to considerable improvements in judicial decisions and thereby enhance societal welfare. Simulations indicate that the use of ML-powered decision support systems may reduce jailing rates by more than 40 percent with no increase in crime rates. Chalfin et al. (2016) outline that machine learning applications can potentially enhance welfare by providing predictions about workers' productivity. They find evidence suggesting that replacing currently used hiring and promotion systems with automated AI systems can be highly effective in increasing organizational efficiency. The authors estimate the benefits of switching to a ML-powered system by replacing the hired (promoted) subjects in the bottom productivity decile with average productive ones and compare the overall productivity of this new distribution with the original one.

In contrast to the limited number of related studies, we do not use a highly specific setting and econometric techniques to approximate causal welfare consequences. We precisely quantify the causal effects of replacing current human-made decisions with those of differently biased AI systems in an abstract setting that mimics the fundamental structure of numerous areas where such machines are already employed today. To the best of our knowledge, we are the first to combine an abstract experimental paradigm with ML applications to produce novel insights into the broad systemic consequences of integrating AI systems into societies. In particular, our unique approach allows us to isolate and showcase downstream effects of algorithmic biases on economic efficiency, which has not been done so far. Finally, our study extends this literature by providing causal evidence on how unrepresentative training data and the ongoing maintenance of algorithms constitute a source of variation in AI systems' potential to benefit social welfare.

The second strand of literature we contribute to are studies on algorithmic fairness, biases, and discrimination. This literature broadly examines how ML algorithms may unintentionally reproduce human stereotypes, biases, and outcomes considered as unfair, e.g., by learning encoded patterns from training data (e.g. Barocas & Selbst, 2016). Over the last couple of years, there has been a steady stream of empirical work documenting how AI systems may impose less favorable treatment on already disadvantaged groups. Examples include racial biases in the recidivism risk assessment (Angwin et al., 2016), predictive policing (Ensign et al., 2017), and health risk assessment (Obermeyer et al., 2019), as well as gender biases in the delivery of ads (Sweeney, 2013; Lambrecht & Tucker, 2019), and in facial recognition tasks (Buolamwini & Gebru, 2018). Because of existing correlations in the data, ML algorithms may even learn to discriminate based on sensitive features, such as gender or race, even if these attributes have been explicitly excluded from the training process (Kleinberg, Ludwig, et al., 2018). Recently, there are also some theoretical contributions outlining that under certain conditions, biased training data may not always be as detrimental to algorithms' performance as one might assume (Cowgill, 2018a; Rambachan & Roth, 2019). Our article contributes to this line of previous work by illustrating how the degree of an AI system's initial bias determines whether or not its use leads to welfare gains or losses. More specifically, we produce causal empirical evidence how non-randomly missing observations in the training data may cause ML algorithms to learn biases and thereby create detrimental consequences for both discriminated and non-discriminated groups.

Finally, we relate to a limited number of articles that are concerned with algorithmic feedback loops. Feedback loops can occur when algorithms shape decisions whose observed outcomes supplement the training data that is fed to the machine in the future, e.g. in the pace of an updating process. Once these outcomes are used as training data to improve existing or develop new algorithms, the contaminated data may reinforce inherent biases (Cowgill & Tucker, 2019). In other words, through feedback loops, algorithms may causally affect the outcomes they are designed to improve. Cowgill (2018b) shows the occurrence of an algorithmic feedback loop in the context of bail decisions. The author uses a regression discontinuity design to show that algorithmic predictions causally affect defendants' re-arrest likelihood - the outcome the algorithm is designed to predict - and thereby endogenously shape the training data used to develop future algorithms. This way, the algorithm's prediction eventually becomes a self-fulfilling prophecy altering the ground truth, in this case for the worse. Even if feedback loops can not change the ground truth, they may cause training data to become increasingly unrepresentative when there exists a selective-labels problem (Lakkaraju et al., 2017). This issue occurs, whenever observations for the training data can only be collected if a decision-maker takes a particular action, e.g., we only learn about a person's creditworthiness if this person receives a loan and thus has the option to pay the loan back at an agreed point in time. Over time, an algorithm may increasingly distort training data by causing a selective enrichment of the data, lowering future predictive performance for underrepresented types (see for example Heckman, 1979). Our results depict that in a stable environment where there is no discrimination, continued updating can create feedback loops that increasingly rectify unrepresentative training data. By repeatedly retraining algorithms on the more and more representative data, even strongly discriminatory AI systems debias themselves over time without exogenous intervention.

Theoretical Framework

To be able to study the causal impact and precisely quantify downstream consequences of letting an AI system instead of a human make decisions in a strategic setting, we choose a controlled, abstract setting. More specifically, we make use of the sequential prisoners' dilemma that reflects the essence of any sequential economic exchange that takes place in the absence of perfect enforcement mechanisms, e.g., because they are prohibitively costly (Fehr & Fischbacher, 2003; Dufwenberg & Kirchsteiger, 2004). Broadly, one may conceive these sequential exchanges as employer-employee exchanges (Akerlof, 1982), principal-agent exchanges (Fehr et al., 1997), or market exchanges (Fehr et al., 1993; Brown et al., 2004).

The basic structure of the game is as follows. A trustor and a trustee are matched in pairs of two. Both players are initially endowed with 10 monetary units (MU). The trustor starts to decide whether to transfer her 10 MU to the trustee - cooperate (C) - or to keep the endowment for herself - defect (D). The trustee observes the trustor's initial decision

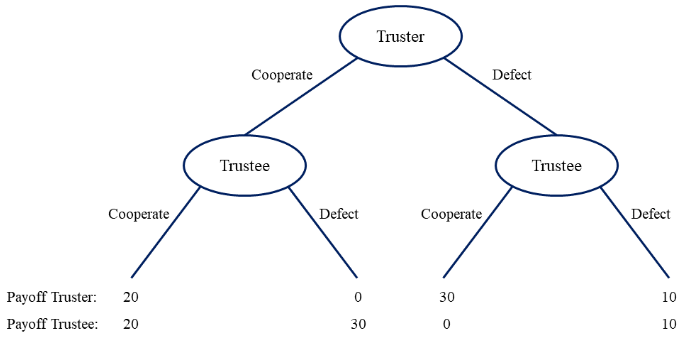


Figure 1. A sequential prisoners' dilemma

and then chooses to cooperate or defect as well. Any MU transferred from one player to the other is doubled (see figure 1 for an illustration). In this structure, two aspects are noteworthy. First, trustors make their initial strategic decision under uncertainty, not knowing how trustees will respond, while trustees possess full information about trustors' choices when deciding. Second, social welfare is maximized in case both players exchange their endowment, i.e., carry out the exchange, while individually there exists a strong incentive for the trustee to cheat and not to behave reciprocally. This is because the trustee's material payoff is maximized when receiving a transfer from the trustor while keeping his initial endowment for himself.

This abstract strategic setting with asymmetric information mimics the fundamental structure of many real-life situations. Examples include, a manager (trustor) who decides upon promoting an employee (trustee) to a supervisor position, who in turn can respond to a promotion by exerting high or low effort, or a loan officer (trustor) managing risk and determining whether to charge high or low interest to an applicant (trustee), who can subsequently choose to repay the credit or default. In these and many more situations, social efficiency dictates the trustor to cooperate (i.e. promote, charge a low interest rate) whenever the trustee, in turn, would reciprocate by cooperating as well (i.e., exert high effort, repay the loan).

Given the lack of information, the trustor needs to assess the likelihood that her counterpart behaves reciprocally. This is where ML algorithms, often as part of a broader information system, come in. Using available information about the trustee, ML algorithms can produce a prediction about the trustee's likelihood to reciprocate initial cooperation by cooperating as well. The prediction as such effectively reduces the asymmetry of information between the trustor and the trustee. Generated algorithmic predictions can then be used, either by humans themselves or another machine, to make an optimal decision. This way, assessments are not based on population averages, intuition, or subjective experience, which is prone to mental errors (see for example Tversky & Kahneman, 1974; Kahneman & Tversky, 1977).

While on an individual level, the use of an ML application to enhance efficiency appears intuitive, the downstream consequences on broad equilibrium outcomes, are more complex and demand a closer analytic contemplation. In the following, we, therefore, derive a simple theoretical framework to illustrate the structure of the setting more formally and show how the deployment of an AI system may affect broad population-wide outcomes.

Assume there is a continuous population of individuals with a total mass normalized to one. This population can be interpreted as a society into which the AI system will be integrated. We model people's engagement in sequential exchanges as follows. The entire population is randomly split up in equal shares of trustors and trustees. Each trustor is randomly matched with one trustee to play a sequential prisoners' dilemma that follows the structure explained before (see figure 1). Let the set of available pure-strategies for trustors be given by $A_1 = \{C, D\}$, where the pure strategies respectively refer to *cooperation* (C) and *defection* (D). The pure-strategy set for trustees is denoted as $A_2 = \{CC, CD, DD, DC\}$. The two letters from left to right respectively indicate a trustee's conditional response to the trustor initially choosing to cooperate and defect. For example, a trustee choosing strategy CC always cooperates independent of the trustor's initial choice, while a trustee choosing CD cooperates if the trustor initially cooperated and defects in case the trustor initially defected.

The material payoff an individual i in the role $k = 1, 2$ receives when choosing strategy $a_{i,k} \in A_k$ depends on the strategy $a_{j,-k} \in A_{-k}$ that the matched opponent j in role $-k$

plays.[1] We denote individual i's payoff as $\pi_i(a_{i,k}, a_{j,-k})$. Following the structure used in our field study, payoffs conditional on the chosen strategies, i.e., game outcomes, are defined as depicted in figure 1.

Let every individual i be described by (θ_i, x_i), with $\theta_i \in \{s, r\}$ denoting individual i's type and x_i being a vector representing this individual's personal characteristics. We assume that s-types, are only concerned with their personal material payoff (*selfish-types*). In the role of a trustee in a one-shot sequential prisoners' dilemma their optimal strategy is always to defect $a^*_{i,2}(s) = DD$. r-types in the role of a trustee, on the other hand, are assumed to behave reciprocally, i.e., $a^*_{i,2}(r) = CD$. The population shares of reciprocal and selfish types are respectively denoted as μ_r and $\mu_s = 1 - \mu_r$.

While a person's type θ_i is private information, we assume that the characteristics x_i of an individual are observed. Notably, we assume that individuals themselves can not infer someone else's type θ_i, and thus trustee behavior, from observing x_i. This could for example be because the relationship is highly non-linear and imposes prohibitively high costs. This implies that there exists a strong asymmetry in information between trustors and trustees.

The observed characteristics x_i, however, can be used as an input for a trained machine learning algorithm that generates a prediction $\hat{\theta}_i \in (0, 1)$ that a person will reciprocate cooperation as a trustee, i.e. that a person is of type $\theta_i = r$. The ML algorithm is trained on a historic data set D comprising a large number of observational pairs (θ, x) drawn from the distribution $P(\theta, x)$. For simplicity we abstract from the estimation problem and denote the trained algorithm as $f_D(x) = \hat{\theta}$. We assume that the trained algorithm is part of a broader AI system that uses the prediction to make utility-maximizing choices on behalf of trustors.

As a representation of individual i's personal preferences, we use a simplified version of the model by Charness and Rabin (2002), which allows for conditional social welfare concerns and has been shown to explain empirical observations of sequential prisoners'

[1] Note: $-k$ reflects that individual j takes on the opposite role of individual i, i.e., $-k = 2$ if $k = 1$, and $-k = 1$ if $k = 2$.

dilemmas extremely well (see Miettinen et al., 2020). We denote an individual i's utility function $U_i(\pi_i, \pi_j, \theta_i)$ as

$$U_i(\pi_i, \pi_j, \theta_i) = \begin{cases} (1 - \rho(\theta_i))\pi_i + \rho(\theta_i)\pi_j & \text{if } \pi_i \geq \pi_j \\ (1 - \sigma(\theta_i))\pi_i + \sigma(\theta_i)\pi_j & \text{if } \pi_i < \pi_j \end{cases}, \tag{1}$$

where $\rho(.)$ and $\sigma(.)$ are type-dependent non-negative parameters with $\sigma(.) \leq \rho(.) < \frac{1}{2}$, indicating the conditional weights individual i puts on her opponent j's material payoff π_j. The AI system that makes decisions on behalf of trustor is individually calibrated to know the stakeholder's utility function.

We model individuals (and the AI system) to act as expected utility maximizers so that the chosen strategy a^* ultimately reflects the solution to the optimization problem

$$a^*_{i,k} = \underset{a_{i,k} \in A_k}{\text{argmax}} \sum_{a_{j,-k} \in A_{-k}} p(a_{j,-k}) \cdot U_i(\pi_i(a_{i,k}, a_{j,-k}), \pi_j(a_{j,-k}, a_{i,k}), \theta_i). \tag{2}$$

$p(a_{j,-k}) \in (0,1)$ denotes individual's i's belief that her opponent j will choose strategy $a_{j,-k} \in A_{-k}$, at the moment when i is making her decision.[2] Given the sequential structure, trustees, when choosing their strategy, observe their opponent's actual choice. Hence, trustees do not face uncertainty about the trustor's behavior and assign the probability of one to the observed choice.

With the outlined maximization problem and the payoff structure defined in 1, we can derive conditions for $\rho(\theta_i))$ and $\sigma(\theta_i))$ for both types $\theta \in (s, r)$. Substituting the payoffs into the utility function, it is trivial to derive that trustees always choose to defect if $\rho(s), \sigma(s) \leq \frac{1}{3}$ while they choose to reciprocate the trustor's strategy if $\rho(r) \geq \frac{1}{3}$ and $\sigma(r) \leq \frac{1}{3}$.

For simplicity, let $\rho(s) = \sigma(s) = \sigma(r) = 0$ and $\rho(r) = \frac{1}{2}$ so that we can rewrite utility

[2] Note: For simplicity we do not allow for type-dependent beliefs.

function (1) as

$$U_i(\pi_i, \pi_j, \theta_i) = \begin{cases} \frac{1}{2}(\pi_i + \pi_j) & \text{if} \quad \pi_i \geq \pi_j, \quad \theta_i = r \\ \pi_i & \text{otherwise} \end{cases} . \tag{3}$$

We now solve the outlined sequential game with imperfect information using perfect Bayesian Nash equilibrium as equilibrium concept. The focus lies on symmetric equilibria in which all individuals possess the same prior concerning the distribution of types in the population and use the same type-dependent strategy. In the following, we, therefore, dispense individual indexation. Equilibrium strategies $a^*(\theta)$ maximize expected utility given a belief about the opponent's strategy p.

The utility function (3) dictates that, independent of their type, it is optimal for trustors to cooperate if $20 \cdot p(C|C) \geq 10$, where $p(C|C)$ denotes trustors common belief that the trustee will cooperate conditional on her own prior cooperation. Since it is common knowledge that there exist only two types in the population, of which merely r-types reciprocate cooperation, we can substitute $p(C|C)$ for the belief $\hat{\mu}_r$ that the trustee is of type r. A trustor, independent of her type, will prefer to cooperate if

$$\hat{\mu}_r \geq \frac{1}{2} \tag{4}$$

This result enables us to derive equilibrium predictions for scenarios where trustors either make the decision on their own or use ML algorithm as an aid to make the choice.

We first consider the case, in which trustors make decisions on their own and do not use an AI system making decisions on their behalf. In this scenario, there are two possible equilibrium outcomes, depending on individuals prior about the share of reciprocal types in the population $\hat{\mu}_r$. Whenever $\hat{\mu}_r < \frac{1}{2}$, no trustor acting as an expected utility maximizer will choose to cooperate. Given trustees' type-dependent optimal strategies $a^*(s) = DD$

and $a^*(s) = CD$, every single game outcome will be mutual defection, i.e., the socially most inefficient outcome. Notably, this outcome constitutes an equilibrium even if the actual share of reciprocal types $\mu_r = 1$, as the inaccurate prior leads to a miscoordination. All proofs can be found in the appendix.

Proposition 1 *Suppose the trustor's belief about the matched trustee's type is $\hat{\mu}_r < \frac{1}{2}$. There exists a unique perfect Bayesian Nash equilibria in which*

$$a^*(s) = a^*(r) = D \tag{5}$$

describe trustors' equilibrium strategies, and

$$a^*(s) = DD \tag{6}$$

$$a^*(r) = CD \tag{7}$$

describe trustees' equilibrium strategies given the belief about the trustors' chosen strategy $p(D) = 1$. In this equilibrium the shares of outcomes on the equilibrium path $\omega\left(a_1^*, a_2^*(a_1^*)\right)$ are given by

$$\omega\left(C, C\right) = 0 \tag{8}$$

$$\omega\left(C, D\right) = 0 \tag{9}$$

$$\omega\left(D, D\right) = 1 \tag{10}$$

Whenever $\hat{\mu}_r \geq \frac{1}{2}$, both types of trustors will always choose to cooperate in order to maximize expected utility. Given the trustee's type-dependent equilibrium strategies, the share of the socially most efficient outcome is at its maximum. This, however, comes at the expense of trustors who are randomly matched with selfish trustees, since they are able to free-ride on trustors' cooperative behavior leaving them with no surplus.

Proposition 2 *Suppose the trustor's belief about the matched trustee's type is* $\hat{\mu}_r \geq \frac{1}{2}$. *There exists a unique perfect Bayesian Nash equilibria in which*

$$a^*(s) = a^*(r) = C \tag{11}$$

describe trustors' equilibrium strategies, and

$$a^*(s) = DD \tag{12}$$

$$a^*(r) = CD \tag{13}$$

describe trustees' equilibrium strategies given the belief about the trustors' chosen strategy $p(C) = 1$. *In this equilibrium the shares of outcomes on the equilibrium path* $\omega\left(a_1^*, a_2^*(a_1^*)\right)$ *are given by*

$$\omega\left(C, C\right) = \mu_r \tag{14}$$

$$\omega\left(C, D\right) = (1 - \mu_r) \tag{15}$$

$$\omega\left(D, D\right) = 0 \tag{16}$$

Overall, propositions 1 and 2 emphasize the role individual beliefs play for coordination and equilibrium selection and how distinct subjective assessments about trustees' propensity to reciprocate cooperation shape social-welfare. Even for large shares of reciprocal types among the population, it is possible that no efficient sequential exchanges take place due to miscoordination under asymmetric information. ML-generated, individual level predictions can ultimately influence the welfare and efficiency of societies in these settings, because they reduce the asymmetry of information between trustors and trustees.

Next, we consider how equilibrium outcomes change when introducing an AI system that makes decisions on behalf of human trustors. The AI system comprises the predictive ML algorithm $f_D(.)$ and the codified preferences of the trustor on whose behalf the system decides. Using the prediction and the preferences, the AI system always chooses the utility-maximizing strategy. As explained before, the ML algorithm uses a trustee's observable characteristics to produce an individual level prediction $f_D(x) = \hat{\theta}$ about the trustee's propensity to reciprocate cooperation. Since the AI system is designed to make an optimal decision given the preferences and the algorithmic prediction, we can simply substitute the common prior for the algorithm's predictions $\hat{\mu}_r = \hat{\theta}$ to model the rule according to which the system decides. According to condition (4), there exists a unique equilibrium in which the AI system will independent of her human stakeholders type cooperate if the individual prediction $\hat{\theta} \geq \frac{1}{2}$ and defect otherwise. Hence, $\frac{1}{2}$ effectively serves as the lower threshold for classifying a trustee as being reciprocal. Together this threshold and the type-dependent probability distribution of algorithmic predictions $q(\hat{\theta}|\theta)$ determine the algorithm's predictive performance and thereby welfare consequences.

Proposition 3 *Suppose an AI system uses an individual-level algorithmic prediction about the matched trustee's type $\hat{\theta}$ to make a utility maximizing choice on behalf of a human trustor. There exists a unique perfect Bayesian Nash equilibrium in which*

$$a^*(s) = a^*(r) = \begin{cases} C & \textit{if} \quad \hat{\theta} \geq \frac{1}{2} \\ D & \textit{otherwise} \end{cases}. \tag{17}$$

describe the AI system's equilibrium strategies, and

$$a^*(s) = DD \tag{18}$$

$$a^*(r) = CD \tag{19}$$

describe trustees' equilibrium strategies given the unity belief about the AI system's chosen strategy. Conditional on the type-dependent probability distribution of algorithmic predictions $q(\hat{\theta}|\theta)$, the shares of outcomes on the equilibrium path $\omega\left(a_1^*, a_2^*(a_1^*)\right)$ are given by

$$\omega\left(C, C\right) = \mu_r \int\limits_{0.5}^{1} q(\hat{\theta}|r) d\hat{\theta} \tag{20}$$

$$\omega\left(C, D\right) = (1 - \mu_r) \int\limits_{0.5}^{1} q(\hat{\theta}|s) d\hat{\theta} \tag{21}$$

$$\omega\left(D, D\right) = (1 - \mu_r) \int\limits_{0}^{0.5} q(\hat{\theta}|s) d\hat{\theta} + \mu_r \int\limits_{0}^{0.5} q(\hat{\theta}|r) d\hat{\theta} \tag{22}$$

Proposition 3 depicts the potential gains but also dangers associated with letting an AI system decide on behalf of human trustors. Whether such a machine ultimately improves social welfare depends on its ability to correctly classify trustees' types. Any ML algorithm $f_D(x)$ that correctly classifies at least one reciprocal-type will prevent the occurrence of the most inefficient equilibrium described under proposition 1. The more reciprocal subjects are correctly classified as such, i.e., the higher $\int_{0.5}^{1} q(\hat{\theta}|r) d\hat{\theta}$, the more socially efficient outcomes occur. In other words, the recall value of the AI systems predictive ML component indicates how useful the system is in terms of facilitating mutual cooperation. Conversely, when the predictive algorithm exhibits a low performance in correctly classifying reciprocal types, it can steer the population into a less efficient state by fostering the occurrence of welfare minimizing outcomes of mutual defection. A strongly biased system which systematically produces overly pessimistic predictions that individuals with a specific characteristic in x_i are reciprocal, i.e., incorrectly low values of $\hat{\theta}$, will thus enhance inefficiencies.

Data collection and simulation design

Our analyses are based on a rich data set that has been collected in a voluntary field study between 2016 and 2019. Participants in this study were first-semester economics students from a large German University. The study was conducted online on *LimeSurvey* and comprises a broad set of survey items on students' demographics, socio-economic background, cognitive abilities, personality traits, and experimental tasks. Overall there are 49 distinct questions.[3] Most important for this work, the study additionally includes an incentivized one-shot sequential prisoners' dilemma. The version used in the field study is identical to the one explained in the previous section (see figure 1 for an overview). Participants' trustee choices were elicited using the strategy method, i.e., they had to indicate whether or not to cooperate for both possible decisions an anonymous trustor could have made. For every participant in the study, we observe the unconditional trustor and both conditional trustee choices. We randomly drew 5 percent of all participants and split them into equal shares of trustors and trustees. Subsequently, we randomly matched them in pairs of two and paid them according to the game outcome that resulted from combining the trustor's unconditional choice with the corresponding conditional decision of the trustee. For each MU earned in the game, chosen participants received 1 Euro. On average participants earned 13.16 Euro.

Overall, we collected 3,624 individual observations that make up our raw data set. The raw data set required considerable preprocessing due to fragmentation. After cleansing the raw data, we are left with 1051 observations.[4] Each observation represents the actual and materially consequential choices of a real person together with information about this person's characteristics. Specifically, each observation comprises this person's trustor decision, both conditional trustee decisions, and answers to 16 questionnaire items. We selected these 16 items because comprehensive empirical testing in regards to feature

[3] We show an overview of all items in the Appendix B in figure 11

[4] Note: To simplify the analyses and facilitate the interpretability of our results, we only use subjects who as trustees always defect, or behave reciprocally. These two types make up for 93% of our usable post-cleansing observations.

engineering and selection revealed that they jointly constitute a set of strong features allowing us to create a high performing ML model. Table 1 shows these items, together with descriptive statistics.

	Item	Scale	Mean	Std. deviation
1.	Big 5: Openness	(0,1)	0.625	0.208
2.	Big 5: Conscientiousness	(0,1)	0.669	0.171
3.	Big 5: Extraversion	(0,1)	0.639	0.221
4.	Big 5: Agreeableness	(0,1)	0.715	0.165
5.	Big 5: Neuroticism	(0,1)	0.522	0.215
6.	Risk aversion	(0,1)	0.542	0.205
7.	Competitiveness score	(0,1)	0.617	0.218
8.	Trust in choice of study	(0,1)	0.711	0.248
9.	Current happiness with choice of study	(0,1)	0.729	0.225
10.	Likelihood of finishing studies	(0,1)	0.822	0.22
11.	Volunteer social year prior to studies	Yes=1, No=0	0.075	0.263
12.	Subject related internship prior to studies	Yes=1, No=0	0.148	0.355
13.	Non-Subject related internship prior to studies	Yes=1, No=0	0.169	0.375
14.	Apprenticeship prior to studies	Yes=1, No=0	0.149	0.356
15.	Foreign language spoken at parental home	Yes=1, No=0	0.287	0.453
16.	Gender	Male=1, Female=0	0.509	0.5

Table 1. Items from field study used as features to train the ML algorithm. Note that we normalized the scale of numeric items 1 to 10.

The objective of this paper is to study individual and population-wide efficiency and welfare effects of integrating different AI systems into human societies. To do so, we use our cleaned data as a basis for distinct simulation exercises. Simulations only differ with respect to the design of the AI system's predictive ML component.

Simulations have the following basic structure, which mirrors our outlined theoretical framework. At the beginning, we randomly split our cleaned data into a training set (75% of observations, i.e., 795 observations) and a population set (25% of observations, i.e., 256 observations). The training set is further preprocessed and then used to train, validate, and test, a ML algorithm that uses a person's 16 characteristics as input features to predict her likelihood to reciprocate cooperation in the role of a trustee. We use an Adaptive Boosted Random Forest method. The forest comprises 100 individual trees with a depth of 5. Adaptive boosting refers to the sequential learning process where each new predictor corrects the predecessor by putting more weight on training instances that were previously underfitted. The Adaptive Boosting method is among the most popular and most powerful ensemble methods (Freund & Schapire, 1997). Our trained algorithms exhibit a high performance on all relevant technical performance measures. Table 4 in the Appendix B shows a performance overview of our algorithms after validation and training on a test set.

Algorithm 1: Sequence of simulation exercises

```
Result: Game outcomes and utilities in sequential prisoners' dilemma games
Cleaning of raw data;
while counter ≤ 10 do
    1. Random partition of cleaned data - 25% population set, 75% training set;
    2. Preparation of training set for training of ML algorithm;
    3. Training, validation, testing of ML algorithm on training set;
    4. Estimation of individual utility functions for subjects in population set;
    while counter ≤ 100 do
        5. Random draw of 50% of individuals in population set;
        6. Random partition of selected individuals in trustors and trustees;
        7. Random matching of trustors and trustees in pairs of two;
        8. Matching of human / AI system trustor decisions with trustees conditional choices,
          determination of game outcomes and utilities.;
        9. Compute diverse performance metrics
    end
end
```

The population set, on the other hand, is used to simulate sequential prisoners' dilemma games. This is done in three steps which are repeated 100 times. First, we randomly select half of the individuals from the population set. Second, the drawn individuals are randomly split in equal shares of trustors and trustees. Third, to determine the outcome and utilities if human trustors make the decisions, we match the original choices of the trustors from the field study with the corresponding conditional response of the trustee. To establish what the AI system does, the trained ML component makes a prediction about the matched trustee's likelihood to reciprocate cooperation. The decision making component subsequently uses this prediction and the trustor's previously estimated utility function[5] to compute whether cooperation or defection yields a larger expected utility. The AI system's decision is the utility-maximizing strategy, which is

[5] We use subjects' trustee decisions from the field study, to estimate individual level parameters of a simplified version of the social preference model by Charness and Rabin (2002), which we explained in detail in the section where we presented our theoretical framework. Utility functions have the form of equation (3), where r-types are subjects from the field study who mirrored the trustor's choices, i.e. are reciprocators.

then matched with the corresponding conditional response of the trustee, to determine outcomes and utilities. Every simulation is replicated 10 times. Overall, each simulation comprises 64,000 distinct games. An overview of this simulation process can be found in the depicted pseudo code 1.

Across different simulations, we vary the AI system's algorithmic bias against female trustees. More specifically, we deliberately manipulate the predictive ML component of AI systems so that it systematically underestimates the probability that a female trustee, relate to a male trustee, will reciprocate cooperation. This is the case even though female individuals in our data set are on average significantly more likely to reciprocate than men (75.4% vs. 68.1%, Wilcoxon rank-sum test $p < 0.000$). We introduce biases by means of imbalancing the training set (in step 2 in pseudo code 1), while holding the overall number of observations fixed. This way we control for the overall amount of training instances. We vary the share of reciprocal examples among women from 0 (no reciprocal women at all) to 0.5 (balanced share of reciprocal and non-reciprocal women) with a step-size of 0.05. With less examples of reciprocal women to learn from, the likelihood of correctly classifying reciprocal (selfish) women will decreases (increase). Male observations in the training data set were perfectly balanced with regards to the label. This is, in the course of preprocessing the data, we ensure that for male observations, there is an equal share of reciprocal and selfish examples in the training set, so that the classification of reciprocal and selfish men works equally well.

We choose the gender attribute as an example of algorithmic discrimination to pin down the consequences of biased systems for two reasons.[6] First, there exists ample scientific and anecdotal evidence showing that algorithmic discrimination against women, e.g. due to previous discriminatory practices encoded in training data, is an actual, considerable societal problem (see for example Sweeney, 2013; Buolamwini & Gebru, 2018; O'Neil, 2018; Lambrecht & Tucker, 2019). Second, male and female participants in our field study

[6] Note: One should understand the use of the gender attribute as a representative example of a broad range of characteristics that algorithms may discriminate on.

exhibit a statistically significant difference in their propensity to reciprocate cooperation in the role of the trustee (respectively 75.4% and 69.1%, χ^2-test: p<0.000). As a consequence, from a technical perspective, the variable gender possesses explanatory power concerning a person's likelihood to behave reciprocally, allowing us to introduce biases in the first place.

Finally, to examine interaction effects between algorithmic biases and continued updating, specifically retraining of the algorithm, as well as algorithmic feedback loops, we deploy a slightly adapted simulation sequence. This sequence differs from the previously explained one (see pseudo code 1) only with regards to the inclusion of two additional steps at the end. In each iteration, after determining game outcomes, the previous training data set is supplemented by trustees (their 16 personal characteristics and their choice when the trustor cooperates) whose matched trustor initially cooperated. Subsequently, we retrain the AI system's predictive ML component on the appended training data. The retrained ML component then makes predictions in the next iteration. With this procedure, the algorithmic prediction endogenously shapes the structure of the training data on which we retrain the algorithm in the next iteration and thus future predictions. As a consequence, our setting allows the occurrence of data-driven feedback loops. An overview of this slightly adapted simulation process can be found in the depicted pseudo-code 2 in the appendix.

Results

The results of our simulation exercises are presented in three parts. First, we examine how well an unbiased AI system, relative to humans, performs in making trustor decisions. Our objective is to answer the question of whether an unbiased AI system can enhance welfare and efficiency on both an individual and population-wide level. These findings serve as a benchmark to map out the potential of an unbiased system. Subsequently, we go over to our main endeavor and outline how results change in case the underlying ML algorithm becomes increasingly biased against women. By doing so we show in detail the role algorithmic biases play regarding AI systems' potential to augment social welfare. Finally, we study to what extent continued learning may, over time, enable a strongly biased ML algorithm to recover itself.

Unbiased AI system

We start our analyses with comparing the performance between human trustors and an unbiased AI system making decisions on behalf of these human trustors.[7] We initially focus on performance differences from the perspective of human stakeholders. We will consider two distinct measures. First, we compare the share of decisions that are optimal from the human trustor's point of view. A decision is individually optimal, i.e., utility-maximizing, whenever (i) the trustor defects in case the trustee would not reciprocate cooperation, or (ii) the trustor cooperates in case the trustee would reciprocate cooperation. Subsequently, we consider differences in average trustor utility across the two scenarios.[8]

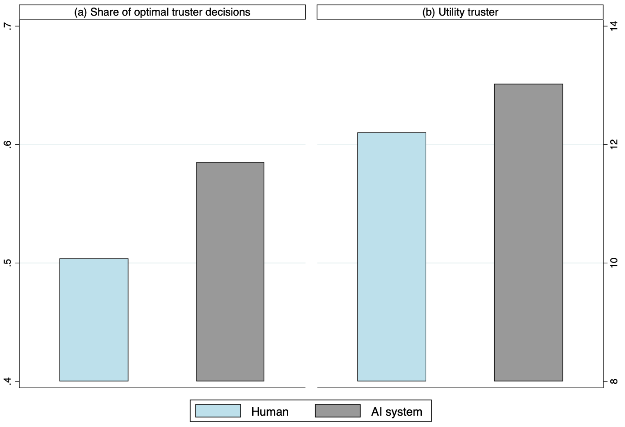


Figure 2. For human and AI system scenarios, panel (a) represents the shares of optimal trustor decisions, while panel (b) shows average trustor utility. trustor utility is depicted in normalized units.

[7] Unbiased refers to the fact that in comparison to other AI systems, we did not intentionally introduce an algorithmic bias in the form of systematically inaccurate predictions against women. A Wilcoxon rank-sum test reveals that the prediction errors between women and men are not significant, despite the large sample size ($p < 0.12$)

[8] Note: Given the structure of the field study, human trustors were not able to observe trustees' characteristics. However, we assume that human trustors, on average, do not exhibit systematic behavioral biases with regards to the 16 characteristics of trustees which the AI system uses as input features to make a prediction. Following this assumption, the law of large numbers suggests that the average human trustor decision should not considerably differ even if trustors would have observed the trustees' characteristics before making their decisions.

Figure 2 depicts the shares of optimal decisions (panel (a)) and the average trustor utility (panel (b)) that human trustors themselves and the AI system on their behalf are able to achieve.

We observe that the AI system significantly outperforms human decision-makers regarding the optimality of individual choices and hence the utility gained for the trustor. On average, a human trustor, depending on the trustee's conditional responses, makes a utility-maximizing choice in 50.3% of the games.[9] The AI system, in contrast, is able to do so in 58.4%. This difference is equal to an increase by 16.1%-points and is statistically highly significant (χ^2-test: $p < 0.000$). The gained utility significantly increases from 12.2 to 13 units (+6.6%) in case the AI system instead of the human makes the initial decisions (Wilcoxon signed-rank test: $p < 0.000$).[10]

Looking at individual choices, we find the AI system to make a different decision than its human stakeholder in 49.9% of the games. Conditional on making a different choice, the AI system improves the stakeholder's position in 29% of the cases. More specifically, the AI system optimally chooses to cooperate (defect) while the human trustor suboptimally defects (cooperates) in 24.3% (4.7%) of the cases. However, in about 21 out of 100 games (20.9%), the system renders the human worse off when making a different decision, either by defecting while cooperation would have been reciprocated (11.4%) or cooperating even though defection would have been the utility-maximizing choice (9.5%). On average, letting the AI system make a different decision than the human trustor pays off. The average utility increases from 11.3 units (human) to 12.9 units (AI system), which is an economically considerable increment of 14%.

Next, we look at the population-wide consequences. Figure 3 shows the shares of overall game outcomes. Panel (a) depicts outcomes for human trustors, while panel (b) illustrates the AI system scenario. CC, DD, and CD respectively refer to outcomes where trustors and trustees both cooperate, where trustors and trustees both defect, and where

[9] Note: If the trustee reciprocates cooperation, the utility-maximizing trustor decision is to initially cooperate; if the trustee does not reciprocate cooperation, the utility-maximizing trustor decision is to initially defect.

[10] Summary statistics on trustor utility are provided in table 2

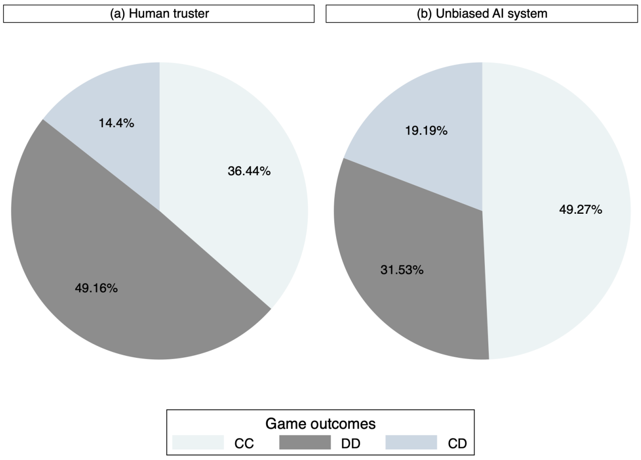


Figure 3. Relative frequencies with which different game outcomes, i.e., mutually cooperative (CC), mutually defective (DD), and free-riding (CD) outcomes, occur. Panel (a) represents results for human decision-makers; panel (b) represents results for an unbiased AI system.

trustors cooperate while trustees defect.

Panel (a) depicts that population-wide, about half of human trustors initially defect (49.2%), causing the socially most inefficient outcome to occur. Notably, in 71.6% of the instances where the trustor initially defects, thereby inevitably evoking the mutually defective outcome, the trustee would have reciprocated initial cooperation, so that the socially optimal outcome could have been reached. On the other hand, human trustors cooperate in 50.8% of the games. This initial cooperation, however, is only reciprocated in 36.4% of the games, while trustees free-ride, and thus prevent the occurrence of the socially most efficient outcome, in 14.4%. Out of all the cases where mutual cooperation would have been feasible, humans are only able to reach this socially most desirable outcome in 50.7%. Overall, these observations illustrate that there is considerable miscoordination and thereby social inefficiencies, raising the question of whether the use of an AI system can overcome this issue.

Observations in panel (b) suggest that the AI system can at least mitigate social inefficiencies, even though not entirely overcome them. The AI system initially defects in 31.5% of the games, which is 17.6% less often relative to humans. While this shows that the most inefficient outcome can be avoided more often, still 70.9% of initial defection remains inefficient, because the trustee would have reciprocated cooperation. Hence, in relative terms, initial machine defection is only marginally more efficient than initial human defection (71.6%). The AI system chooses to cooperate in 68.5% of the games. Thereby, the socially most efficient outcome is reached about half of the time (49.3%). Compared to the human benchmark, this is an increase by 12.9%, or 35.4%-points, and thus economically highly considerable. Mutual cooperation can be implemented in 68.8% of all the cases where it is possible (human benchmark: 50.7%).

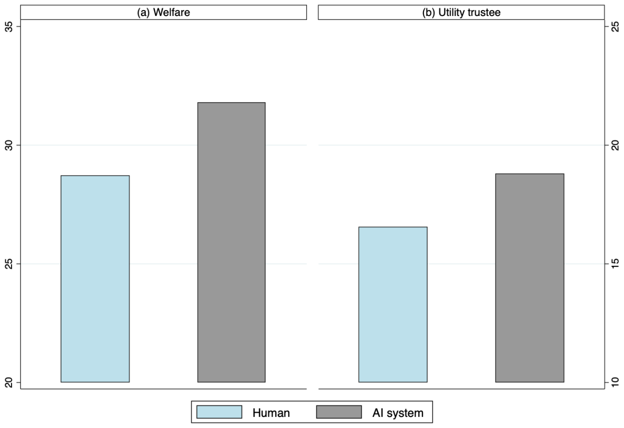


Figure 4. For human and AI system scenarios, panel (a) represents average welfare, while panel (b) shows average trustee utility.

Naturally, the AI system's higher performance in reaching the socially most efficient outcome translates into enhanced social welfare. Figure 4 shows the average population welfare and trustee utility for the two scenarios. We observe a statistically significant and economically relevant increase in welfare from 28.7 to 31.8 units (+10.7 %) when the AI system makes trustor decisions instead of humans themselves (Wilcoxon signed rank test: $p < 0.000$). Note that if we only consider cases where the AI system's decision has led to a Pareto improvement, i.e., exclude games where the human reaches the mutually defective outcome while the AI ends up in the free-riding outcome, the increase is equal to 6.9% (Wilcoxon signed rank test: $p < 0.000$). Hence, even when excluding the outcomes where social welfare increases at the expense of the trustor, the overall increase in social welfare remains economically and statistically significant.

	Welfare	Utility trustor	Utility trustee
Human trustor	28.7	12.2	16.5
	(9.162)	(6.777)	(7.177)
AI system	31.8	13	18.8
	(8.813)	(7.708)	(7.014)

Table 2. Summary statistics on welfare and utility levels. Displayed measures are mean values. Standard errors are reported in parentheses. The number of observations equals N = 64000.

Depicted measures further show that the observed increase in social welfare does not only result from higher trustor utility, but also enhanced trustee utility. Trustees' average utility increases by 13.9% (from 16.5 to 18.8 units; Wilcoxon signed rank test: $p < 0.000$). In comparison to the trustors, this increase is relatively larger (13.9% vs. 6.6%). This finding can be explained by the observation that non-reciprocal trustees can free-ride on initial cooperation more often, which benefits trustees but reduces the utility of trustors to 0. In other words, due to their informational advantage from moving second, all trustees benefit from the higher rate of initial cooperation, even if it is not optimal from the trustor's perspective.

Result 1 *An unbiased AI system significantly outperforms its human stakeholder when making decisions in the role of the trustor. On average, the share of individually optimal decisions increases by 8.1%, leading to an increase in utility by 6.6%. On a population-wide level, the AI system increases social welfare by 10.7 % (6.9% when considering Pareto improvements only). The increase in welfare results from both, increases in trustors' and trustees' utility.*

Biased AI systems

After we have seen that an unbiased AI system outperforms human trustors to the benefit of the entire population, we now go over to study how algorithmic biases affect these results. When we talk about an AI system exhibiting an algorithmic bias, we refer to the predictive ML component of an AI system producing systematically incorrect predictions for a specific group of individuals, which leads to wrongfully unequal treatment.

We intentionally introduce an algorithmic bias of the AI system against women by training ML algorithms that, ceteris paribus, estimate women to be less likely to reciprocate cooperation than men, despite them being more likely to do so. AI systems comprising these biased ML algorithms are more likely to defect when interacting with a female trustee, compared to male.

We create biases by imbalancing the training set. In imbalanced sets, the share of non-reciprocal female observations exceeds the fraction of reciprocal ones for a fixed level of female observations. The data available to train the ML algorithm is therefore a non-representative subsample for women. We vary the share of reciprocal examples among women in the training set from 0 (no reciprocal women at all) to 0.5 (fully balanced shares of reciprocal and non-reciprocal women) with a step-size of 0.05. The balanced case is the benchmark that we analyzed in the previous section.

To see that we successfully bias the ML algorithm by imbalancing the training set, consider table 3. The table depicts the average predicted probabilities that women and men cooperate, conditional on the degree of imbalance. The table shows that the ML algorithm increasingly underestimates the likelihood that women in the player set reciprocate cooperation when the relative share of reciprocal female examples decreases. For men, the average predicted probabilities are about the same across different degrees of imbalance. Wilcoxon rank-sum tests reveal that, except for the 50% case ($p < 0.12$), the average predictive errors are significantly different for women and men ($p < 0.000$ for all other cases). The ML algorithm thus learns an incorrect representation of women's trustee behavior, while the representation for men is more precise so that the system produces systematically less favorable predictions for women.

Probability of	True	Share of reciprocal examples among female observations in the training set										
being reciprocal	measure	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Women	0.75 (0.435)	0.03 (0.107)	0.15 (0.257)	0.25 (0.314)	0.36 (0.344)	0.42 (0.364)	0.5 (.0358)	0.56 (0.357)	0.59 (0.35)	0.64 (0.339)	0.64 (0.33)	0.69 (0.318)
Men	0.69 (0.462)	0.58 (0.389)	0.58 (0.375)	0.58 (0.371)	0.58 (0.362)	0.58 (0.356)	58 (0.352)	0.59 (0.35)	0.6 (0.354)	0.61 (0.342)	0.63 (0.339)	0.62 (0.34)

Table 3. The true share of reciprocal individuals in the player set and the mean predicted probabilities of different ML algorithms are displayed. Standard errors are reported in parentheses.

Given that we have successfully introduced algorithmic discrimination in our framework, we now examine to what extent our previous results depend on the degree of this algorithmic bias against women. We start examining how biases influence the AI system's performance relative to human decision-makers from the perspective of the trustor. Subsequently, we outline population-wide efficiency and welfare ramifications.

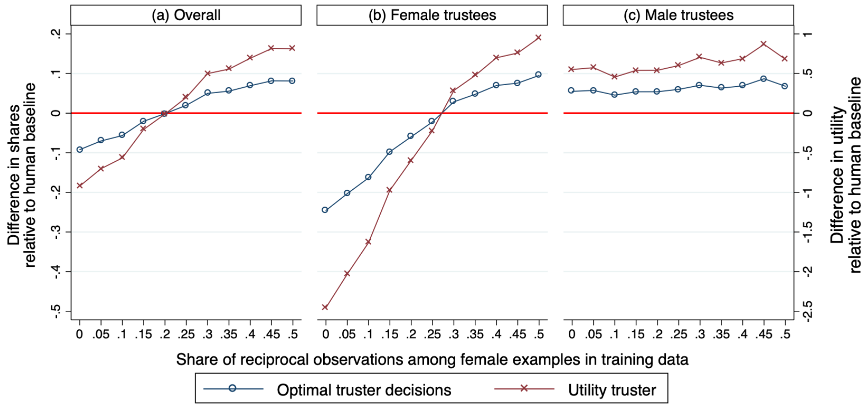


Figure 5. Differences in shares of optimal trustor decisions and trustor utility relative to human baseline, conditional on the degree of algorithmic bias against women. From left to right panels show results for (a) the entire sample of games, (b) the sumsample of games with female trustees, and (c) the subsample of game with male trustees.

Figure 5 depicts the performance of AI systems, relative to human decision-makers, conditional on the degree of bias. Depicted plots show the difference in the share of optimal trustor decisions and the average utility of trustors. Positive values on the Y-axes indicates that the machine outperforms human trustors, while negative ones indicate the reverse. Panels (a), (b), and (c) respectively show results for all games, the subsample of games where trustees are women, and the subsample of game where trustees are men.

The figure portrays that an AI system's capability to outperform its human stakeholder in making trustor decisions critically depends on the presence of an algorithmic bias against specific groups. The more inaccurate algorithmic prediction against women, the worse are the decisions made by the AI system. Whenever the share of reciprocal examples among females in the training set is smaller than 20%, the share of utility-maximizing trustor decisions made by the AI system is smaller than the measure for human decision-makers. Only for a share of 25% of reciprocal females, or more, an AI system, on average, makes better choices than human trustors. Naturally, worse decisions translate into lower utility, so that a trustor who lets a strongly biased AI system make decisions on her behalf is worse off in comparison to deciding herself. The system with the most potent bias leads to a considerable reduction in trustors' average utility of 7% (from 12.2 to 11.3).

Intuitively, the negative effects of algorithmic biases for trustors are driven by instances where the trustee is a woman (see panel (b)) since the predictive performance is low for this group of individuals. In our framework, if the most biased system (0% of reciprocal observations among females) makes trustor decisions instead of humans, the average utility of trustors drops by 2.5 units when the trustee is a woman. This is equivalent to a decrease of 20% and economically substantial. With regards to male trustees (panel (c)), the AI system outperforms and increases the utility for their human stakeholders, independent of the degree of the bias. Notably, the relative performance of an AI system when a trustee is a man even increases slightly when the training set becomes more balanced for women. A possible explanation for this observation is that a more balanced subset of female data points implies that the training data at large also becomes more balanced.

Overall, these observations illustrate that it is in the interest of trustors that the AI system that decides on their behalf is unbiased, in particular, if they are likely to interact with the algorithmically discriminated group. This is because the bias implies that the AI system's performance in making optimal decisions for the trustor is systematically worse when the trustee belongs to the algorithmically discriminated group.

Result 2 *Strongly biased AI systems perform significantly worse than their human stakeholders, when making decision as trustors. The more biased the system, the worse off is the human stakeholder.*

Next, consider how biases against women affect the population as a whole. Figures 6 and 7 respectively illustrate population-wide effects in terms of how the occurrence of specific outcomes and welfare as well as trustee utility differs from the human benchmark. We show results for all games (panel (a)), the subsample of games where trustees are women (panel (b)), and the subsample of games where trustees are men (panel (c)).

Figure 12 in Appendix B gives an overview of absolute shares of game outcomes for all different AI systems.

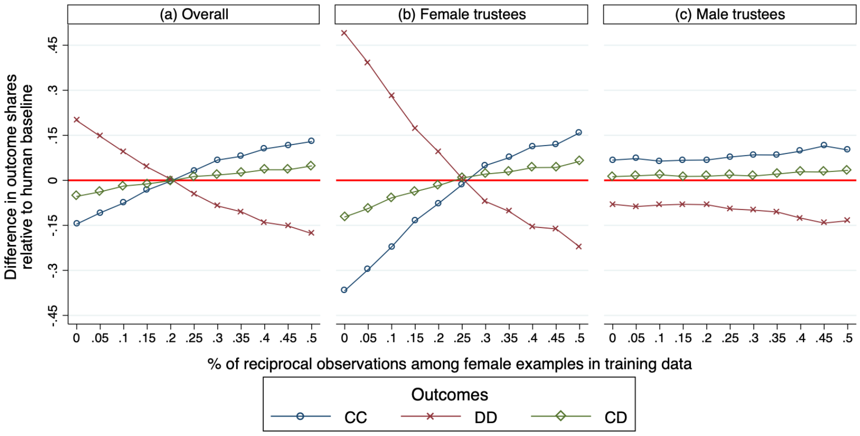


Figure 6. Differences in relative frequencies with which specific outcomes occur relative to human baseline, conditional on the degree of algorithmic bias against women. From left to right panels show results for (a) the entire sample of games, (b) the sumsample of games with female trustees, and (c) the subsample of game with male trustees.

Both figures emphasize the detrimental population-wide consequences that the use of biased AI systems may entail. The more biased a system is, the more it increases (decreases) the occurrence of the socially most efficient (inefficient) outcome relative to a human decision-maker (see figure 7). In our setting, compared to the human trustor, the most biased system reaches the mutually defective outcome 19.9% more often, while the occurrence of mutual cooperation drops by 14.6%. Until the part of female training instances comprises at least 20% of reciprocal examples, the use of an AI system steers the entire population into a less efficient state compared to human decision making. These negative ramifications are entirely driven by games where trustees are female. In cases where the trustee is male, the mutually cooperative outcomes increase while the mutually defective ones decrease independent of the training data distortions. Hence, algorithmic biases create considerable differences in game outcomes, based on gender. In the most biased case, the AI system only cooperates in 1.5% of the cases where it would have been optimal to do so in case the trustee is a woman. In contrast, this AI system does so in 60.3% when a trustee is a man.

Due to highly biased systems' inefficiently low cooperation with female trustees, social

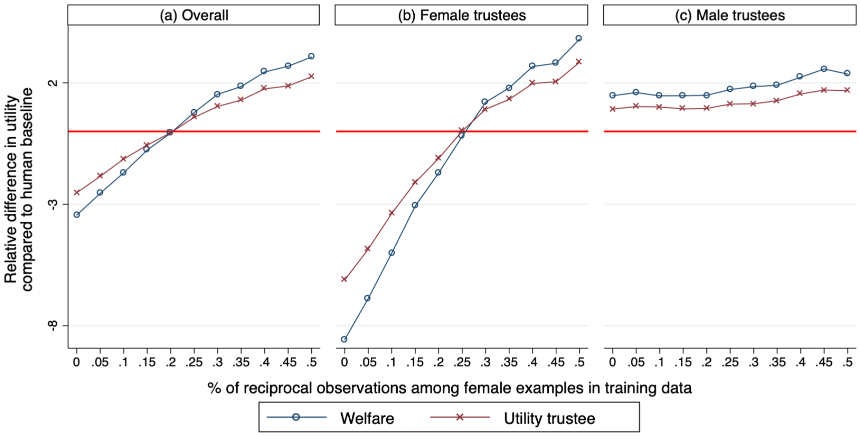


Figure 7. Differences in welfare and trustee utility relative to human baseline, conditional on the degree of algorithmic bias against women. From left to right panels show results for (a) the entire sample of games, (b) the sumsample of games with female trustees, and (c) the subsample of game with male trustees.

welfare decreases substantially. Concerning the most biased system, welfare subsides by 11.8% (from 28.7 to 25.3 units). Highlighting the severely unequal treatment in the most biased case, it is the group of female trustees who bear the brunt of the welfare loss since their average drop in utility is even larger than the mean loss experienced by the trustor (-37.8% vs. -20%). In contrast, the utility of male trustees grows by 6% in this scenario. This is particularly concerning, considering that algorithmic biases are often the digital continuation of historical discrimination and less favorable treatment of specific groups, that has been encoded into training data. Panel (c) in figure 6.2 further suggests that male trustees also have an interest in interacting with an AI system that does not discriminate against women, since they benefit more when the AI system is less biased. The utility of male trustees increases by 10.2% when an unbiased system instead of a human makes the trustor decision; when interacting with the most biased one only by 4% (Wilcoxon rank-sum test: $p < 0.000$).

Result 3 *Strongly biased AI systems may steer entire populations into undesirable and socially highly inefficient states. The discriminated group bears the brunt of the harm. The potential to augment social welfare is thus inextricably linked to a system's resilience not to inherit discriminatory behavior in the training process.*

Finally, there is one firmly important issue we want to stress. In the setting we consider, the true label of a trustee, i.e., whether this person reciprocates cooperation or not, is only observed in case the trustor initially cooperates. Initial defection always leads to a defective response of the trustee, which does not provide useful information about this person being a reciprocator or not. This selective labels issue (Lakkaraju et al., 2017), reflects the fundamental structure of a multitude of real-life situations in which algorithms are used to automate or augment decisions. Examples include patrolling decisions of the police (Ensign et al., 2017), bank officers issuing loans (Huang et al., 2007), and judges making bail decisions (Kleinberg, Lakkaraju, et al., 2018), to name only a few.

In our study, we are in an unusual position to observe a trustee's response even for trustor choices that did not actually happen. As a consequence, we are able to compute precise performance metrics. In real-life scenarios, however, one naturally does not observe the accuracy of a prediction that evokes the decision where no label is produced, e.g. one does not know whether a negative prediction about a person's creditworthiness is accurate if the predictions leads to the decision not to issue a loan. The measurement and assessment of an algorithm's performance is thus limited to the selectively generated outcomes, which may lead to incorrect conclusions.

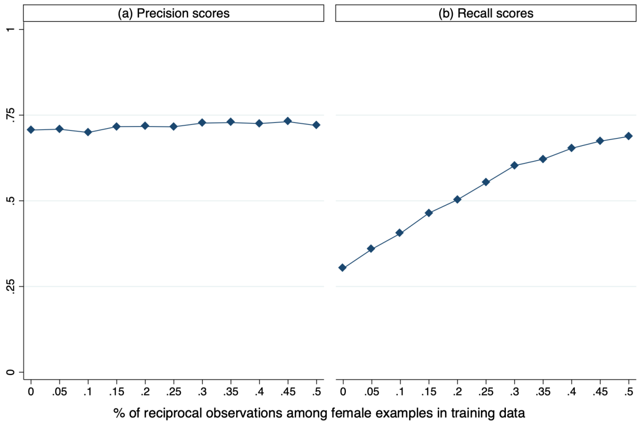


Figure 8. Predictive performance of the ML algorithm conditional on the degree of bias. Panel (a) depicts the precision metric. Panel (b) the recall metric

To illustrate this issue, consider figure 8 which portrays performance metrics for AI systems conditional on their inherent algorithmic bias. Panel (a) depicts the share of optimal trustor decisions given that the system cooperated (i.e. the precision score), which is the measure that is available in real-life situations. Panel (b), on the other hand, shows the share of optimal decisions given that cooperation would have been reciprocated (i.e. the recall score), which is generally not available in real-life scenarios.[11]

The figure depicts an alarming pattern. Independent of the inherent algorithmic bias, and thus of the efficiency and welfare consequences, the precision metric indicates that about 71% of the decisions to cooperate are correct (see panel (a)). This conveys the impression that all AI systems perform equally well. Even the most biased, and welfare reducing, AI system may be incorrectly assessed as performing reasonably well, if one bases the evaluation on this metric. Panel (b), on the other hand, paints a more accurate picture of the AI systems' performance. It shows that the recall score is sensitive to the algorithm's bias. The more pronounced the bias, and thereby the negative efficiency and welfare consequences, the lower is the value of this performance metric. For instance, instead of indicating that the most biased system performs about as well as the unbiased one (respective precision scores: 0.71 vs. 0.72), the recall score shows a considerably lower performance for the most biased system (recall scores: 0.3 vs. 0.69). Unfortunately, it is not possible to retrieve the recall measure in cases where labels are generated selectively, only the precision score. This emphasizes the importance of a careful interpretation and assessment of available performance metrics on AI systems, especially in environments where the problem of the selective labels likely occurs. If one uses these measures as a basis to decide about the continued or enhanced employment of these machines, there could be detrimental society-wide ramifications without decision-makers even knowing that a more efficient outcome would have been feasible.

Result 4 *In an environment of selective labels, the accurate evaluation of algorithmic performance is difficult and prone to be misleading.*

[11] The accuracy of the AI systems decisions, which is unobservable in real-life, is depicted in figure 5 as the share of optimal decisions.

Algorithmic biases and continued learning

So far our results emphasize that AI systems are a two-edged sword. On the one hand, we map out how an unbiased AI system can create substantial welfare gains for the entire population. This is mainly because the AI system effectively reduces asymmetric information and correctly chooses to cooperate more often than its human stakeholders do. On the other hand, we provide controlled evidence that the integration of biased systems may not only limit the positive consequences but reverse them to the negative and create considerable population-wide welfare losses. The main reason for that appears to be that based on the systematically incorrect prediction, biased AI systems seize cooperation. These observations imply that in order to maximize the potential benefits of AI systems for societies, it is important to further our understanding of how to counteract algorithmic biases.

We, therefore, devote the final part of our analyses to studying how algorithmic biases endogenously change if they continue to learn within an environment, where the originally learned biases are no longer present. The notion of why this may be the case is as follows. ML algorithms learn from data that is assumed to be drawn from a fixed, unknown distribution. When algorithms learned to make systematically incorrect predictions for unseen out-of-sample examples, it is from a technical perspective because the distributions from which the training and out-of-sample examples are drawn from differ fundamentally. If we interpret this difference as being the result of a change in a non-stationary environment, algorithmic biases, at least in terms of systematically incorrect predictions, can be interpreted as an inherent concept drift, i.e., a fundamental change in the representation to be learned (Widmer & Kubat, 1996). In the domain of learning in non-stationary environments, the literature has argued that continued learning may be a natural remedy to deal with concept drifts by adapting learned representations dynamically over time (see for example Jordan & Mitchell, 2015; Elwell & Polikar, 2011).

Following this notion, we study the development of algorithmic biases over time, when the ML component of our AI system is repeatedly retrained using training data that is supplemented by previous game outcomes from the population. We consider 100 rounds of play where we retrain the ML algorithm using the original training set supplemented by the game outcomes of all previous periods. This setting mirrors a scenario, where a fixed population of individuals interacts with each other over a certain period. Note that continued learning in our context technically implies that the original training set is increasingly supplemented by a limited number of distinct observations from the population set. As a consequence, the predictive ML algorithm will likely overfit after some periods. The point of the analyses, however, is to document whether continued learning on a fixed population that systematically differs from the original training set, can mitigate algorithmic discrimination over time. Therefore, the issue of overfitting is of secondary importance to our endeavor.

To ensure a better overview, we will focus on three AI systems, that differ with regards to the intended bias we initially introduce through distorting the original training data. We consider (i) an unbiased AI system where female examples are balanced with regards to the labels, (ii) an intermediately biased AI system where the share of reciprocal examples among female observations equals 20%[12] , and (iii) a strongly biased AI system where there are no reciprocal female examples in the original training data. At this point, it is important to emphasize the selective labels setting. Given the structure of the game and the predictive ML algorithm, observed game outcomes can only supplement the training data, in case the AI system cooperates. As a consequence, a continuous extension of the training data with selective observations also bears the risk of further distorting the data used to (re)train the predictive algorithm, so that existing biases are maintained or even reinforced via feedback loops (Cowgill & Tucker, 2019).

Since the detrimental population-wide consequences of employing biased algorithms can ultimately be traced back to a systematically incorrect prediction about women's likelihood to reciprocate cooperation, we look at the development of predictions by the

[12] Note: We choose 20% as intermediately biased since previous analyses revealed that for this share of reciprocal examples among females, the AI system leads to almost the same outcomes as in the human benchmark.

ML algorithm over time.

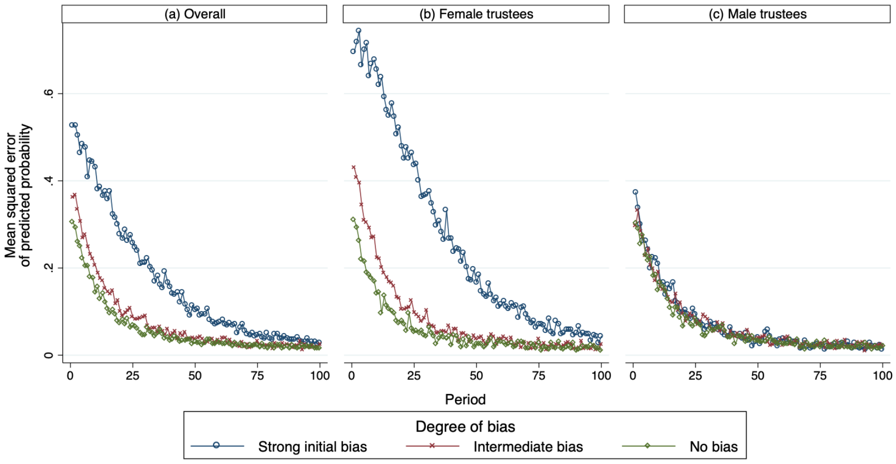


Figure 9. Mean squared errors of predicted probabilities are depicted over time. From left to right panels show results for (a) the entire sample of games, (b) the sumsample of games with female trustees, and (c) the subsample of games with male trustees.

Figure 9 shows the development of the mean squared error of the predicted probability that a trustee is a reciprocator over time under continued learning. We display results for the overall sample of games (panel (a)) and subsamples of games with female and male trustees (respectively panel (b) and (c)). Illustrated results indicate that continued learning in our setting, at least to some extent, provides a remedy for algorithmic biases over time. By using the response and characteristics of trustees against whom the AI system cooperated as additional observations to supplement training data and retrain the algorithm, the predictive performance of all three algorithms increases substantially over time. Even for the most biased algorithm, the mean squared error for the entire sample decreases from 0.52 to 0.26 after 25 rounds of play (see panel (a)). After 50 periods, the error further dropped to 0.1. This decrease is driven by both, improved performance when the trustee is a woman and a man. Notably, while the predictive error for men is still smaller than for women (0.04 vs. 0.17), the difference has decreased from initially 0.33 (0.37 vs. 0.7) to 0.13. With regards to the intermediately biased algorithm, the initial difference in the performance between men and women even vanishes entirely (from 0.3 vs. 0.43 to 0.04 vs. 0.04). The displayed results further suggest that the degree and speed with which continued learning can mitigate algorithmic biases does depend on the extent of the original bias. The mean squared error curve in panel (b) for the intermediately biased algorithm is found to be steeper than the one for the strongly biased algorithm. Corresponding curves in panel (c) are virtually identical. This suggests that the algorithm with the intermediate bias unlearns systematically incorrect predictions for women, in favor of more accurate ones, faster than the algorithm with the strongest bias. One explanation, corroborated by our data, is that the less biased system initially cooperates more with female trustees and thus creates larger amounts of additional training data which helps to improve the predictive performance.

In general, it appears that feedback loops drive the observed self-correction process. By increasingly supplementing original training data with observations from the population set, the original differences in the training and population sets disappear. Retraining the ML algorithm on more and more representative training data, helps increasing its predictive performance. Thereby the AI system correctly cooperates more often, which in turn leads to an accelerating enrichment of the training data with new, representative observations. Given that the most biased system initially barely cooperates with female trustees (only in about 1% of the cases), it seems that even a few additional observations can, after some time, invoke the self-correcting feedback loop.

Overall, these observations emphasize that continued learning may lead to considerable increases in predictive performance, which are associated with a decrease in algorithmic biases. The improved performance and mitigated biases naturally translate into positive efficiency and welfare consequences (see figures 13, 14, and 15 in the appendix).

Result 5 *Continued learning can improve ML algorithms predictive performance over time. Supplementing the training data with affected outcomes and retraining the algorithm mitigates algorithmic biases.*

Discussion and Conclusion

With the paper at hand, we contribute to discussions about the broad consequences of integrating AI systems into human societies. We use a game-theoretic setting that provides us the necessary control over potential confounds and allows us to observe counterfactual outcomes of choices. More specifically, we make use of the sequential prisoners' dilemma paradigm, a setting that mirrors the fundamental structure of a multitude of real-life situations. Our objective is to produce causal evidence on how algorithmic discrimination influences AI systems' potential to augment individual and population-wide welfare.

Our results show that the employment AI systems can significantly improve economic efficiency and social welfare on an individual and a population-wide level. The change in efficiency and welfare associated with letting AI systems instead of humans decide, however, depends on the extent of inherent algorithmic biases. In our setting, AI systems that make systematically incorrect choices when interacting with females, can cause considerable efficiency losses and decrease social welfare, especially for the discriminated groups. Considering that algorithmic biases often originate from historic discrimination that is encoded in data, biased AI systems entail the risk of maintaining and, depending on their scope of application, scaling discriminatory practices. This is particularly concerning given that inherent algorithmic biases are frequently hard to detect so that discrimination may already have been institutionalized and led to considerable social problems for the disadvantaged group. In that sense, our results emphasize the importance to ensure that broadly employed AI systems work accurately for all groups. To this end, it is vital to identify adequate performance metrics. However, as shown, this can be particularly difficult in selective labels settings, where algorithmic performance can only be measured on a highly endogenous subsample of outcomes, so that even algorithms that do very poorly convey a false impression of performing well. This emphasizes the danger that algorithmic discrimination, with its negative ramifications, remains hidden over a long period.

Additional findings in our paper also show a silver lining in this regard. In particular, our analyses suggest that continued learning can provide a remedy to systematically inaccurate ML behavior. In that regard, our insights indicate the superiority of continuously learning AI systems over static ones in domains where there is a strong likelihood that predictive algorithms are originally trained on data suffering from non-randomly missing observations through past sample-selection. Static algorithms that are not improved over time and will always exhibit a low performance with regards to discriminated groups. Algorithms that continue to learn may autonomously improve their predictive performance for underrepresented groups over time due to inherent, data-driven feedback loops. Against this background, organizations may be well advised to implement a process ensuring the continued collection of new training examples and updating of employed AI systems.

Finally, we hope to inspire future research on algorithmic feedback loops and their interaction with algorithmic discrimination. From a policy maker's perspective, it is important to understand how interventions intended to ban human discriminatory practices may interact with biased, continuously learning AI systems in the long run. Especially when algorithmic discrimination is hard to detect and thus likely to remain unaddressed explicitly, it is vital to have insights into dynamic relations between regulation and AI systems so that organizational and political reforms can be better informed.

Appendix A: Proofs

Let $U_i(a_i, a_j)$ denote the utility of trustor i given that the trustor chooses strategy $a_i \in (C, D)$ and the assigned trustee chooses to respond $a_j \in (C, D)$ conditional on observing a_i. There are two types $\theta \in (r, s)$ - reciprocal (r) and selfish (s) - whose preferences are given by

$$U_i(\pi_i, \pi_j, \theta_i) = \begin{cases} \frac{1}{2}\left(\pi_i + \pi_j\right) & \text{if} \quad \pi_i \geq \pi_j, \quad \theta_i = r \\ \pi_i & \text{otherwise} \end{cases}. \tag{23}$$

π_i and π_j respectively describe material payoffs earned by the trustor and the trustee. r- and s-types' optimal pure strategies in the role of the trustee are respectively given by $a^*(r) = (CD)$ and $a^*(s) = (DD)$. The two letters from left to right respectively indicate a trustee's conditional response to the trustor initially choosing to cooperate and defect.

$\hat{\mu}_r$ describes trustors' common prior that an assigned trustee is a reciprocal type. Given the population only comprises reciprocal (r) and selfish types (s), initial cooperation is the utility-maximizing decision for trustor i iff

$$\hat{\mu}_r \cdot U_i(C, C) + (1 - \hat{\mu}_r) \cdot U_i(C, D) \geq U_i(D, D). \tag{24}$$

The game structure and payoffs given a certain outcome equal the following structure:

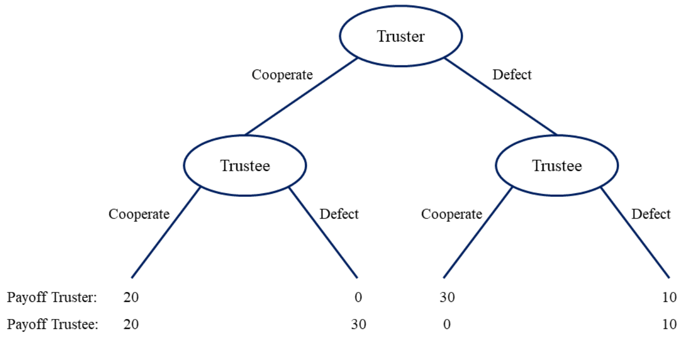


Figure 10. A sequential prisoners' dilemma

Given the depicted payoff structure, it holds for both types that $U_i(C, C) = 20, U_i(C, D) =$

0 and $U_i(D,D) = 10$. As a consequence, we can rewrite condition (24) as

$$\hat{\mu}_r \cdot 20 + (1 - \hat{\mu}_r) \cdot 0 \geq 10. \tag{25}$$

Proof Proposition 1:

Whenever a trustor's personal belief is equal to $\hat{\mu}_r < \frac{1}{2}$, condition 25 can never be satisfied since $\hat{\mu}_r \cdot 20 < 10 \quad \forall \hat{\mu}_r < \frac{1}{2}$. As a consequence, trustors always choose to defect. Given r- and s-types' optimal decisions in the role of the trustee, initial defection is always responded by defection, so that mutual defection is the ultimate outcome.

Proof Proposition 2:

Whenever a trustor's personal belief is equal to $\hat{\mu}_r \geq \frac{1}{2}$, condition 25 is always satisfied since $\hat{\mu}_r \cdot 20 \geq 10 \quad \forall \hat{\mu}_r \geq \frac{1}{2}$. As a consequence, trustors always choose to cooperate. Given r- and s-types' optimal decisions in the role of the trustee, initial cooperation is reciprocated by r-types and responded with defection by s-types. Given the population shares of r- and s-types μ_r and $\mu_s = 1 - \mu_r$, the outcome of mutual cooperation occurs μ_r times of the cases, while the free-riding outcome occurs $1 - \mu_r$ times of the cases.

Proof Proposition 3:

Let an AI system comprise the predictive ML algorithm $f_D(.)$ and the codified preferences of the trustor on whose behalf the system decides. $f_D(x) = \hat{\theta} \in (0,1)$ denotes an individual level prediction that a trustee is of type r. The AI system always chooses the strategy that maximizes the trustor's utility. Hence, the AI system chooses to cooperate iff

$$\hat{\theta} \cdot 20 + (1 - \hat{\theta}) \cdot 0 \geq 10 \tag{26}$$

which is the case whenever $\hat{\theta} \geq \frac{1}{2}$. Let $q(\hat{\theta}|\theta)$ be the type-dependent probability distribution of algorithmic predictions. Given this distribution, the AI system eventually (i) cooperates given the trustee is an r-type with probability of $\int_{0.5}^{1} q(\hat{\theta}|r)d\hat{\theta}$, (ii) defects given the trustee is an r-type with probability of $1 - \int_{0.5}^{1} q(\hat{\theta}|r)d\hat{\theta} = \int_{0}^{0.5} q(\hat{\theta}|r)d\hat{\theta}$, (iii) cooperates given the trustee is an s-type with probability of $\int_{0.5}^{1} q(\hat{\theta}|s)d\hat{\theta}$, and (iv) defects given the trustee is an s-type with probability of $1 - \int_{0.5}^{1} q(\hat{\theta}|s)d\hat{\theta} = \int_{0}^{0.5} q(\hat{\theta}|s)d\hat{\theta}$. Depending on the actual population shares of r-types μ_r and s-types $1 - \mu_r = \mu_s$, the outcome of (i) mutual cooperation occurs $\mu_r \int_{0.5}^{1} q(\hat{\theta}|r)d\hat{\theta}$ times of the cases, (ii) mutual defection occurs $(1 - \mu_r) \int_{0}^{0.5} q(\hat{\theta}|s)d\hat{\theta} + \mu_r \int_{0}^{0.5} q(\hat{\theta}|r)d\hat{\theta}$ times of the cases, and (iii) free-riding occurs $(1 - \mu_r) \int_{0.5}^{1} q(\hat{\theta}|s)d\hat{\theta}$ of the cases.

Appendix B: Supplementary material

Questions on family background

Personal background

How far do you live from your parents?

Please select only one of the following answers:

- I live at my parents
- 1-10 KM away
- 11-50 KM away
- 51-150 KM away
- More than 150 KM away

Have you, due to your studies, changed your place of residence?

Please select only one of the following answers:

- Yes
- No

How many siblings do you have?

Please enter your answers below:

- Younger siblings
- Older siblings

Please indicate with which hand you prefer to perform the following activities:

	Always right	Mostly right	Both hands	Mostly left	Always lfet
Write					
Throw					

Tooth brushing
Holding a spoon

What languages do you speak at home? (multiple answers are possible)

Please select all applicable answers:

- German
- Another language

What is the highest professional qualification of your parents? (Please indicate the highest educational level in each case)

	Father	Mother
University		
University of applied science		
Technical college (former GDR)		
Technician or master craftsman examination		
Apprenticeship		
No educational background		
Unknown		

How do you finance yourself? (multiple answers are possible)

Please select all applicable answers:

- My parents support me financially
- BAföG
- Scholarship
- Job as student assistant (Hiwi) at the university
- Job as a tutor at the university
- Job outside the university
- Other

Questions about the school

School education

At which type of school did you get your university entrance qualification?

Please select only one of the following answers:

- Grammar School
- Comprehensive school
- Vocational school
- Other

After how many school years did you receive your university entrance qualification?

Please select only one of the following answers:

- After less than 12 years
- After 12 years
- After 13 years
- After more than 13 years

In which federal state did you acquire your university entrance qualification?

Please select only one of the following answers:

- Baden-Württemberg
- Bavaria
- Berlin
- Brandenburg
- Bremen
- Hamburg
- Hesse
- Mecklenburg-Western Pomerania
- Lower Saxony
- North Rhine-Westphalia
- Rhineland-Palatinate
- Saarland
- Saxony
- Saxony-Anhalt
- Schleswig-Holstein
- Thuringia
- Other

Which of the following subjects did you take at school in the upper school and what grades (between 1.0 and 4.0) did you have in these subjects in your Abitur certificate?

Please select a maximum of 4 answers.

Please select the appropriate items and write a comment:

- German
- English
- Mathematics
- Physics

Which of these subjects did you take as advanced courses at school?

Please select all applicable answers:

- German
- English
- Math
- Physics
- None of these subjects

Questions on the choice of study subject

I chose my present course of study because...

On a scale from 1 (completely correct) to 6 (completely incorrect) please indicate the accuracy of the following statements.

I chose my present course of study because...

- it particularly interested me and I wanted to
- it corresponds to my inclinations and talents.
- as a graduate of this course of studies I expect particularly good earning and employment opportunities.
- I didn't know what else to do
- I was influenced in my decision by my family / friends

Is your current course of study your dream study?

Please select only one of the following answers:

- Yes
- No

On a scale from 1 (completely sure) to 5 (completely unsure) please indicate the accuracy of the following statements.

- How confident are you in your choice of study?
- How satisfied are you today with your choice of study?
- How certain are you that you will complete your studies?
- How certain are you that you will complete your studies at this university?

Did you do one or more of the following activities before starting your current studies?

Please select all applicable answers:

- Internship related to your field of study

- Internship not related to the field of study
- Training
- Completed studies
- Aborted studies
- Voluntary social year, German Armed Forces, Federal Voluntary Service etc.
- Other:

Questions about studies

Study

How many semesters do you estimate you will need in total until you graduate from your current course?

Only numbers may be entered in this field.

Please enter your answer here:

What are your plans for the time after graduation from your current course of study?

Please select only one of the following answers:

- Begin a further study (e.g. Master's degree)
- go to work
- Other

Based on my grade point average, I expect to belong to...

Please select only one of the following answers:

- ... the top 10% of my class.
- ... the top 11-20% of my year.
- ... the top 21 - 30% of my year of study.
- ... the top 31 - 40% of my year of study.
- ... the top 41 - 50% of my year of study.
- ... the top 51 - 60% of my year of study.
- ... the top 61 - 70% of my year of study.
- ... the top 71 - 80% of my year of study.
- ... the top 81 - 90% of my year of study.
- ... the top 90 - 100% of my year of study.

How important is it to you to maintain your grade point average in your studies or even improve?

Please select only one of the following answers:

- Very important
- Pretty important
- Indifferent
- Rather unimportant
- Very unimportant

How many hours a week do you think you should invest in your studies?

Only numbers may be entered in this field.

Please enter your answer here:

How many hours do you think you will actually invest in your studies each week?

Only numbers may be entered in this field.

Please enter your answer here:

How many hours a week do you currently invest in your studies?

Only numbers may be entered in this field.

Please enter your answer here:

Do you believe that your future earnings will depend on your final grade in your studies?

Please select only one of the following answers:

- Completely correct
- Fully applicable
- Applies
- Applies less
- Not applicable

Risk, Impatience, TC & Narcissism

We would like to ask you to answer the following truthfully. There are no "real" or "wrong answers."

How do you personally assess yourself? Are you generally a person willing to take risks or do you try to avoid risks? Please answer using the following scale, where the value 0 means: "Not willing to take risks at all", and the value 10: "Very willing to take risks". With the values in between you can grade your assessment. Please select the appropriate answer:

- 1

- 2
- 3
- 4
- 5
- 6
- 7

How do you personally assess yourself? Are you generally a person who is impatient or who is always very patient?

Please answer using the following scale, where the value 0 means "very impatient" and the value 10 means "very patient". With the values in between you can grade your assessment. Please select the appropriate answer:

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

To what extent do you agree with the following statement: "I'm a narcissist." (Note: A narcissist is selfish, self-centered, vain.)? Please answer using the following scale, where a value of 1 means "do not agree at all" and a value of 7 means "agree completely". With the values in between you can grade your assessment. Please select the appropriate answer:

- 1
- 2
- 3
- 4
- 5
- 6
- 7

How would you assess yourself in the context of the following statements? Please answer using the following scale, where 1 means "do not agree at all" and 7 means "agree completely". The values in between allow you to grade your assessment. Please select the appropriate answer:

- I like to find myself in situations where I am in competition with others.
- It is important to me to be better than others.
- I think it is important to win at work and in games.

- I exert more effort when competing with others.

Big 5 and Grit

In the list below are different characteristics a person can have. It is likely that some characteristics will apply fully to you personally and others not at all. For others, you may be undecided. Please answer using the following scale:

A score of one means you are not applicable at all.

The value 7 means: fully applicable.

With the values between 1 and 7 you can grade your opinion.

I am someone who...

Please select the appropriate answer:

- works thoroughly
- is communicative, talkative
- is sometimes a little rough on others
- is original, brings in new ideas
- is often worried
- pardon
- is rather lazy
- can come out of itself,
- is sociable
- appreciates artistic, aesthetic experiences
- easily nervous
- Tasks completed effectively and efficiently
- is reserved
- is considerate and friendly with others
- has a vivid imagination, imagination
- is relaxed, can handle stress well

To what extent do the following statements apply to you personally? There are no right or wrong answers here. Please select only one answer in each line.

Please answer using the following scale:

A value of one means they do not apply at all.

The value 5 means: completely correct.

With the values between 1 and 5 you can grade your opinion. Please select only one answer in each line.

- I often set myself a goal, but then decide later to pursue a different goal.
- New ideas and projects sometimes keep me away from previous ones.
- I am interested in something new every few months.
- My interests change from year to year.
- I was once obsessed with a project or idea for a short time, but later I lost interest.
- I find it difficult to stay focused on projects if they last several months.
- I have worked for years towards a goal that I have achieved.
- To overcome important challenges, I also overcome setbacks.
- Everything that I start, I also finish.
- I am not discouraged by setbacks.
- I am a hard working person.
- I am a diligent person.

Trust and Reciprocity

For the following decision situation, another survey participant will be assigned to you randomly. You and this other person make different decisions, which then result in your payout and the payout of the other person. At the beginning you and the other person will each receive 10 Euros from us. You have the following two options to choose from:

Option A: You keep your 10 Euros.

Option B: You give your 10 euros to the other person. The 10 Euros are doubled, i.e. the other person receives 20 Euros.

The other person also has these two options to choose from. Hence, there are four possible outcomes, depending on how you and the other person decide:

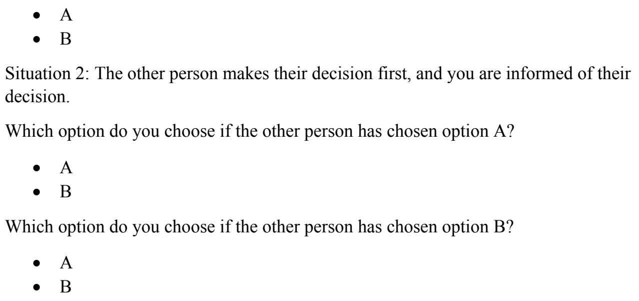
If you and the other person both choose option A, you will both end up with 10 Euros each.

If you and the other person both choose option B, both of you will each have 20 euros.

If you choose option A and the other person chooses option B, you will have 30 euros and the other person 0 euros. And vice versa, if you choose option B and the other person chooses option A, you have 0 euros and the other person has 30 euros. In the following two situations, please decide whether you would rather choose option A or option B. The situations differ in whether you or the other person makes their decision first.

Situation 1: You decide first and the other person is informed of your decision.

Which option do you choose?



- A
- B

Situation 2: The other person makes their decision first, and you are informed of their decision.

Which option do you choose if the other person has chosen option A?

- A
- B

Which option do you choose if the other person has chosen option B?

- A
- B

Figure 11. Translation of field study question items.

	Share of reciprocal examples among female observations in training set										
Performance measure	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Accuracy	0.84	0.82	0.81	0.76	0.76	0.79	0.75	0.74	0.76	0.76	0.75
Precision (Reciprocal)	0.71	0.71	0.74	0.67	0.71	0.74	0.69	0.72	0.76	0.77	0.77
Precision (Selfish)	0.88	0.85	0.83	0.80	0.78	0.80	0.78	0.76	0.77	0.75	0.73
Recall (Reciprocal)	0.61	0.59	0.57	0.57	0.55	0.62	0.63	0.64	0.70	0.71	0.73
Recall (Selfish)	0.91	0.91	0.91	0.86	0.87	0.87	0.82	0.82	0.82	0.80	0.77

Table 4. Algorithmic performance conditional on the share of reciprocal examples among female observations in the training set. We show precision and recall metrics for both types of predictions.

Algorithm 2: Sequence of simulation exercises with continued learning

```
Result: Game outcomes and utilities in sequential prisoners' dilemma games
Cleaning of raw data;
while $counter \leq 10$ do
    1. Random partition of cleaned data - 25% population set, 75% training set;
    2. Preparation of training set for training of ML algorithm;
    3. Training, validation, testing of ML algorithm on training set;
    4. Estimation of individual utility functions for subjects in population set;
    while $counter \leq 100$ do
        5. Random draw of 50% of individuals in population set;
        6. Random partition of selected individuals in trustors and trustees;
        7. Random matching of trustors and trustees in pairs of two;
        8. Matching of human / AI system trustor decisions with trustees conditional choices,
          determination of game outcomes and utilities.;
        9. Compute diverse performance metrics;
        10. Append training data by trustees whose matched trustor cooperated;
        11. Retrain the AI system's ML algorithm on the appended training set
    end
end
```

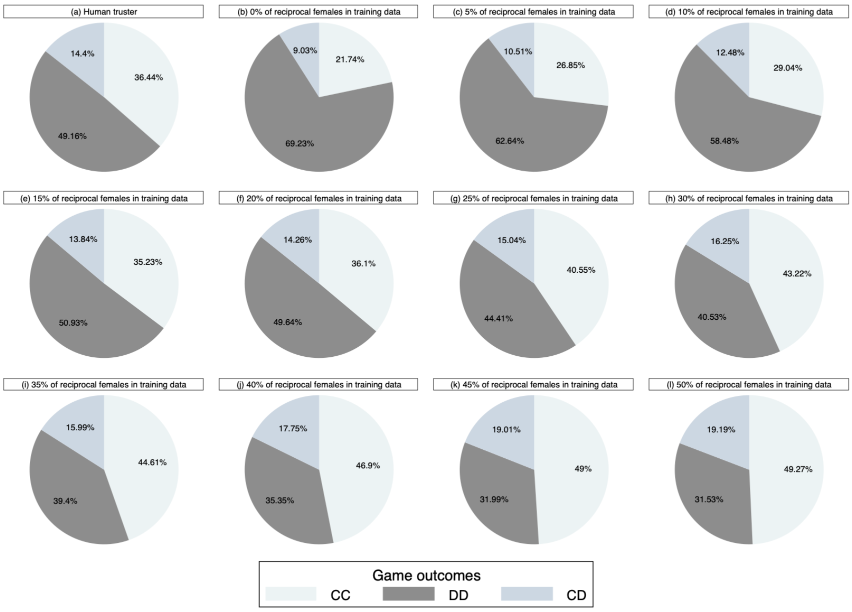


Figure 12. Relative frequencies with which different game outcomes, i.e., mutually cooperative (CC), mutually defective (DD), and free-riding (CD) outcomes, occur. Panel (a) represents results for human decision-makers; panel (b) represents results for an unbiased AI system.

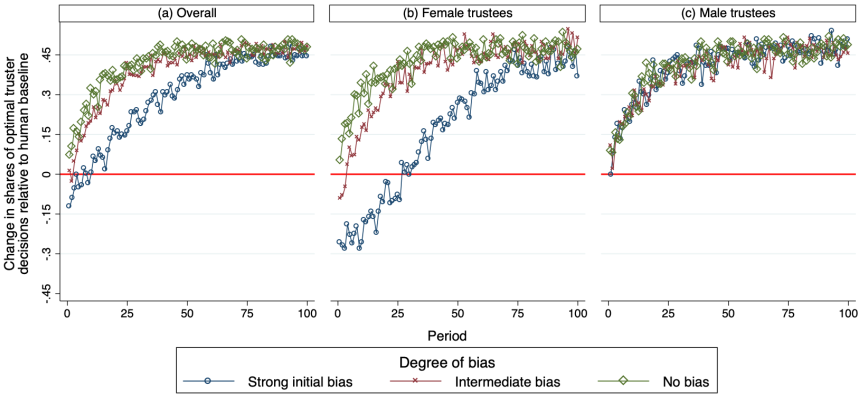


Figure 13. Shares of optimal trustor decisions, conditional on the degree of algorithmic bias against women. From left to right panels show results for (a) the entire sample of games, (b) the sumsample of games with female trustees, and (c) the subsample of game with male trustees.

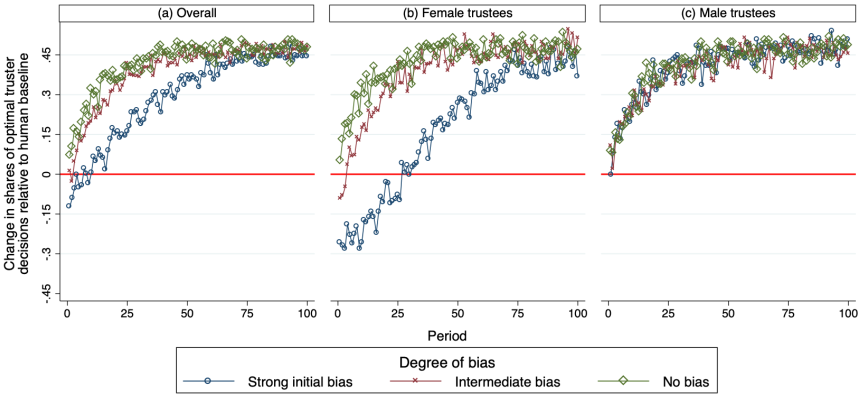


Figure 14. Frequencies with which certain outcomes occur, conditional on the degree of algorithmic bias against women. From left to right panels show results for (a) the entire sample of games, (b) the sumsample of games with female trustees, and (c) the subsample of game with male trustees.

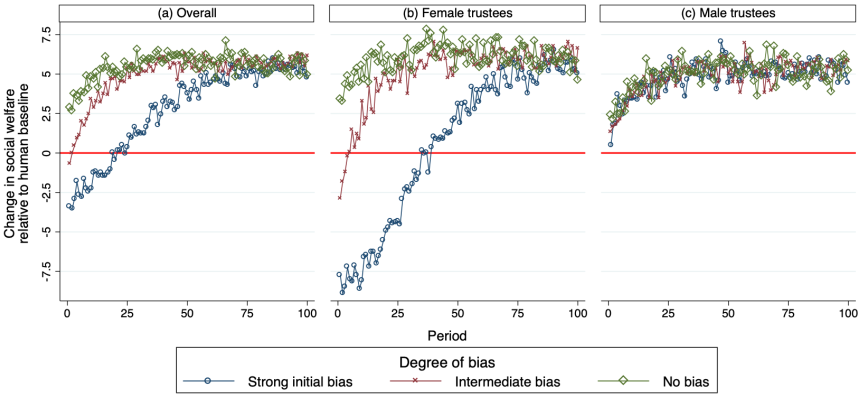


Figure 15. Welfare and trustee utility, conditional on the degree of algorithmic bias against women. From left to right panels show results for (a) the entire sample of games, (b) the sumsample of games with female trustees, and (c) the subsample of game with male trustees.

References

Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management*, *8*(2), 937–953.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, *47*, 1–6.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, *97*(4), 543–569.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*, 2016.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507–547). University of Chicago Press.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602–613.

Brown, M., Falk, A., & Fehr, E. (2004). Relational contracts and the nature of market interactions. *Econometrica*, *72*(3), 747–780.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Chaboud, A. P., Chiquoine, B., Hjalmarsson, E., & Vega, C. (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, *69*(5), 2045–2084.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124–27.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Cowgill, B. (2018a). Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University*, *29*.

Cowgill, B. (2018b). *The impact of algorithms on judicial discretion: Evidence from regression discontinuities* (Tech. Rep.). Technical Report. Working paper.

Cowgill, B., & Tucker, C. E. (2019). Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298.

Elwell, R., & Polikar, R. (2011). Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, *22*(10), 1517–1531.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791.

Fehr, E., Gächter, S., & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica*, 833–860.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? an experimental investigation. *The Quarterly Journal of Economics*, *108*(2), 437–459.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119–139.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 153–161.

Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, *66*(1), 1–33.

Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2010). Matching and sorting in online dating. *American Economic Review*, *100*(1), 130–63.

Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, *133*(2), 765–800.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kahneman, D., & Tversky, A. (1977). *Intuitive prediction: Biases and corrective procedures* (Tech. Rep.). Decisions and Designs Inc Mclean Va.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In *Aea papers and proceedings* (Vol. 108, pp. 22–27).

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 275–284).

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

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, *7*(1), 29.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 60–68.

Miettinen, T., Kosfeld, M., Fehr, E., & Weibull, J. (2020). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. *Journal of Economic Behavior & Organization*, *173*, 1–25.

Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, *107*(5), 476–80.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Nilson. (2016). *Nilson report.* Retrieved 2020-07-15, from `https://nilsonreport.com/upload/content_promo/The_Nilson_Report_10-17-2016.pdf` (Accessed: 15.07.2020)

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

O'Neil, C. (2018). *Amazon's gender-biased algorithm is not alone.* Retrieved 2020-07-29, from `https://www.bloomberg.com/opinion/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone` (Accessed: 29.07.2020)

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … others (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

Rambachan, A., & Roth, J. (2019). Bias in, bias out? evaluating the folk wisdom. *arXiv preprint arXiv:1909.08518*.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, *11*(3), 10–29.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wang, H., Xu, Q., & Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PloS one*, *10*(2), e0117844.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23*(1), 69–101.

LEIBNIZ INSTITUTE
SAFE
Sustainable Architecture
for Finance in Europe

Recent Issues

No. 286	Andreass Hackethal, Michael Kirchler, Christine Laudenbach, Michael Razen, Annika Weber	On the (Ir)Relevance of Monetary Incentives in Risk Preference Elicitation Experiments
No. 285	Elena Carletti, Tommaso Oliviero, Marco Pagano, Loriana Pelizzon, Marti G. Subrahmanyam	The COVID-19 Shock and Equity Shortfall: Firm-Level Evidence from Italy
No. 284	Monica Billio, Michele Costola, Iva Hristova, Carmelo Latino, Loriana Pelizzon	Inside the ESG Ratings: (Dis)agreement and Performance
No. 283	Jannis Bischof, Christian Laux, Christian Leuz	Accounting for Financial Stability: Bank Disclosure and Loss Recognition in the Financial Crisis
No. 282	Daniel Munevar, Grygoriy Pustovit	Back to the Future: A Sovereign Debt Standstill Mechanism IMF Article VIII, Section 2 (b)
No. 281	Kevin Bauer	How did we do? The Impact of Relative Performance Feedback on Intergroup Hostilities
No. 280	Konstantin Bräuer, Andreas Hackethal, Tobin Hanspal	Consuming Dividends
No. 279	Tobin Hanspal, Annika Weber, Johannes Wohlfart	Exposure to the COVID-19 Stock Market Crash and its Effect on Household Expectations
No. 278	Sandra Eckert	EU Agencies in Banking and Energy Between Institutional and Policy Centralisation
No. 277	Dominique M. Lammer, Tobin Hanspal, Andreas Hackethal	Who Are the Bitcoin Investors? Evidence from Indirect Cryptocurrency Investments
No. 276	Massimiliano Caporin, Loriana Pelizzon, Alberto Plazzi	Does Monetary Policy Impact International Market Co-Movements?
No. 275	Loriana Pelizzon, Max Riedel, Zorka Simon, Marti Subrahmanyam	Collateral Eligibility of Corporate Debt in the Eurosystem
No. 274	Christopher Busch, Alexander Ludwig	Higher-Order Income Risk Over the Business Cycle

Leibniz Institute for Financial Research SAFE | www.safe-frankfurt.de | info@safe-frankfurt.de