The Economic Consequences of Algorithmic Discrimination: Theory and Empirical Evidence

Kevin Bauer
Leibniz Institute for Financial Research SAFE, Theodor-W.-Adorno-Platz 3, D-60323 Frankfurt am Main, Germany, Bauer@safe.uni-frankfurt.de

Nicolas Pfeuffer
Chair of Information Systems and Information Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Pfeuffer@wiwi.uni-frankfurt.de

Benjamin M. Abdel-Karim
Chair of Information Systems and Information Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Abdel-Karim@wiwi.uni-frankfurt.de

Oliver Hinz
Chair of Information Systems and Information Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Hinz@wiwi.uni-frankfurt.de

Michael Kosfeld
Chair of Organization and Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Germany, Kosfeld@econ.uni-frankfurt.de

Using a novel theoretical framework and data from a comprehensive field study we conducted over a period of three years, we outline the causal effects of algorithmic discrimination on economic efficiency and social welfare in a strategic setting under uncertainty. We combine economic, game-theoretic, and applied machine learning concepts allowing us to overcome the central challenge of missing counterfactuals, which generally impedes showcasing economic downstream consequences of algorithmic discrimination. Using our framework and unique data, we provide both theoretical and empirical evidence on the consequences of algorithmic discrimination. Our unique empirical setting allows us to precisely quantify efficiency and welfare ramifications relative to an ideal world where there are no information asymmetries. Our results emphasize that Artificial Intelligence systems' capabilities in overcoming information asymmetries and thereby enhancing welfare negatively depend on the degree of inherent algorithmic discrimination against specific groups in the population. This relation is particularly concerning in selective-labels environments where outcomes are only observed if decision-makers take a particular action so that the data is selectively labeled. The reason is that commonly used technical performance metrics like the precision measure can be highly deceptive and lead to wrong conclusions. Finally, our results depict that continued learning, by creating feedback loops, can help remedy algorithmic discrimination and associated negative effects over time.

Key words: algorithmic discrimination, social welfare, economics, game theory, feedback loops, artificial intelligence, machine learning

1. 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 can be used to generate accurate predictions about a variable of interest (label) using available data (features) not included in the training data (Mullainathan and 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 (e.g. Bhattacharyya et al. 2011, Adewumi and 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 et al. 2018a), 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 is a growing stream of evidence indicating how the broad use of algorithms can impose less favorable treatment to already disadvantaged groups creating societal tensions and potential welfare losses, a phenomenon frequently referred to as algorithmic discrimination (Sweeney 2013, Ensign et al. 2017, Obermeyer et al. 2019, Lambrecht and 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, especially discriminatory ones, may scale into population-wide consequences.

With the paper at hand, we intend to contribute to this necessity by studying the economic ramifications of algorithmic discrimination. Similar to related studies, we broadly consider algorithmic discrimination as the algorithmic production of outputs that are inaccurate for a specific group of individuals thereby leading to unfair and disadvantageous economic (and social) outcomes for these individuals, as compared to individuals not belonging to this group (see for example Adomavicius and Yang 2019). We intend to outline and precisely quantify how algorithmic discrimination in a strategic setting under asymmetric information can create considerable efficiency and welfare losses. To this end, we pursue a strategy combining economic, game-theoretic, and applied machine learning paradigms. This allows us to produce both complementary theoretical and empirical evidence on the matter at hand. Specifically, we first examine the relation between economic outcomes and algorithmic discrimination by deriving a novel game-theoretic framework capturing the fundamental informational and incentive structure of a wide range of sequential economic transactions. Subsequently, we put the fundamental model implications to an empirical test using a unique and considerably rich data set that we collected over a period of three years in a large field study.

The central challenge when it comes to an empirical evaluation of economic ramifications of employing discriminatory AI systems lies mainly in assessing whether the AI system's actual decisions are better than the alternative, i.e., choices the system does not make. Put differently, there is a lack of counterfactual observations in real-world field settings. As a consequence, it is almost impossible to assess the welfare ramifications of employing discriminatory AI systems. For instance, if a discriminatory system chose not to hire an applicant while a non-discriminatory system would have done so, it is not possible to measure which decision would have been better simply because there is no data on the applicant's performance had he been hired.

In the empirical part of our paper, we overcome the problem of missing counterfactuals by making use of data that we collected in a controlled and incentivized field study that we conducted over three years. Participants in the field study engaged in an incentivized sequential prisoners' dilemma, i.e., we paid them according to their actions to elicit their revealed preferences, from which we use a specific subset of participants. The basic structure of the game we use in the paper at hand 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. If the trustor decides not to transfer his endowment, the game ends and both the trustor and trustee earn 10 monetary units. If the trustor, however, decides to make a transfer, 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 from one player to another, the monetary units sent from one player to the other are doubled. While abstract, this setting mirrors the incentive and informational structure of any one-shot sequential economic exchange that takes place in the absence of perfect enforcement mechanisms. Specifically, (i) there is an information asymmetry between first and second moving parties, and (ii) there is a conflict between individual and collective interests to engage in, or reciprocate, a transfer. With this structure, the game that we study reflects, for instance, anonymous financial market transactions (Fehr et al. 1993, Brown et al. 2004) and one-shot principal-agent exchanges (Fehr et al. 1997). We elicited field study participants' behavior using the strategy method, i.e., in the role of the trustee, participants make decisions conditional on the decisions of the trustor. Hence, the strategy method gives us the unique opportunity to observe the consequences of counterfactual choices that real people in the role of the truster did not make. Participants in our study also answered a broad set of survey items on demographics, socio-economic background, cognitive abilities, and personality traits.

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 to reciprocate 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. We then conduct comprehensive simulations where the AI system, in the role of the trustor, repeatedly plays on behalf of field study participants against other participants from the field study. Both the AI system's human stakeholders and trustees stem from a population of field study participants that we do not use to build the AI system, i.e., a real out-of-sample population. To determine game outcomes, we match the AI system's decisions with these trustees' actual choices from the field study. Put differently, we use real-life, consequential choices actual

people previously made and not simulated, artificial decisions. Hence, our results depict the social welfare and economic efficiency that the out-of-sample population of field study participants would reach if they were to interact with the AI system. We quantify the level of economic efficiency and social welfare relative to the first-best scenario where there is no uncertainty and human trustors perfectly anticipate responses of trustees if they initially transfer their endowment, i.e., where there is no need to produce and harness ML based predictions. This is only possible since we observe counterfactual trustee decisions.

We start estimating efficiency and welfare consequences if the AI system's ML component does not unfairly produce inaccurate predictions for any specific group of individuals. Subsequently, we show how these results change when the ML algorithm makes inaccurately low predictions about women's likelihood to reciprocate a transfer (while the predictions for men are accurate) so that initial transfers occur significantly less often when the system plays against women compared to men. We induce discriminatory outputs by using non-representative, imbalanced training data, a problem that is highly relevant in practice. Finally, inspired by notions from papers that study ML in non-stationary environments (Elwell and Polikar 2011), we examine whether continued learning - the ongoing updating of ML models using newly collected training examples - can help to counteract originally learned discrimination over time.

There are three main insights from our study. First, in line with a theoretical framework we derive, we produce causal empirical evidence that AI systems' capabilities to enable economic efficiency and social welfare (on both an individual and a population-wide level) critically depends on the absence of inherent algorithmic discrimination against specific subgroups. The more an AI system discriminates, 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 the level of discrimination. Notably, even the group against which the AI system does not discriminate is better off if the predictive ML component does not discriminate. 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, misleading technical performance measures for the employment of the machines. Independent of algorithmic discrimination and welfare consequences, the selectively observed outcomes suggest that all AI systems perform equally well concerning technical performance metrics. This is the case even though strongly discriminating systems create considerable welfare losses that we can only observe in our study because we have access to counterfactuals which are usually not accessible in business practice and most 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 steadily monitoring employed

AI systems and also consulting non-technical performance measures to assess their efficacy. Additionally, this result calls for the development of novel, human-centric monitoring and performance assessment strategies. Finally, we demonstrate that continued learning in a stable environment where there is no discrimination can, at least to some extent, repair originally discriminatory algorithms. The introduction of an updating apparatus creates feedback effects through which initial 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 finding 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 our 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 economic efficiency and welfare. We explain details of the field study we conducted over the last couple of years, the structure of the data, and the simulation exercises in section 4. Section 5 presents our results. Finally, section 6 discusses findings and concludes.

2. Related Literature and Research Hypotheses

Our study aims to document and precisely measure causal efficiency and welfare consequences from letting discriminating AI systems make strategic decisions under uncertainty on behalf of human stakeholders. To this end, we choose an intentionally abstract sequential exchange setting enabling us to observe the ramifications of counterfactual choices (consequences of choices that have not actually been made). This provides us a unique opportunity to isolate the economic ramifications of introducing discriminatory systems on both individual and population-wide levels relative to a first-best scenario in which there do not exist information asymmetries. In our setting, AI systems exhibit different degrees of discriminatory behavior 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 algorithmic discrimination and its consequences. 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 and Selbst 2016). Over the last couple of years, there has been a growing stream of empirical work indicating 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 and Tucker 2019), and in facial recognition tasks (Buolamwini and Gebru 2018). Due to 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 et al. 2018b). 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 and Roth 2019). In another recent study, Adomavicius and Yang (2019) discuss a novel, fairness-aware pipeline for augmented decision-making systems emphasizing that overcoming algorithmic discrimination does not merely require a technical resolution of underlying algorithms but also an understanding and alignment of human behavior and economic incentives. Their work emphasizes the central and strategic that human decision-makers take over when it comes to correcting algorithmic discrimination. Our article contributes to this line of previous work by illustrating how the degree of an AI system's discrimination against a subgroup in the population causally determines whether or not its use leads to welfare gains or losses. More specifically, we produce causal evidence of how non-randomly missing observations in the training data may cause ML algorithms to learn discriminatory practices and thereby create detrimental welfare and efficiency consequences for both discriminated and non-discriminated groups. 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. Instead, we combine simulations on a rich, real-life data set, an abstract experimental paradigm, and game theory with ML concepts to outline how discriminatory AI systems may entail systemic consequences.

The second literature we relate to is a limited number of articles 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 discrimination (Cowgill and 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 (e.g. 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.

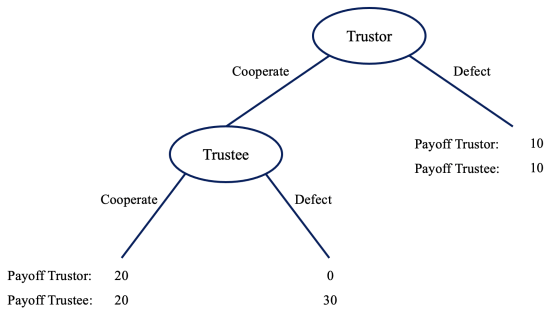
Finally, we also relate to articles that broadly assess the consequences of employing AI systems to augment or automate human decision-making. In the context of medical diagnosing, Mullainathan and Obermeyer (2017) argue that the use of predictive ML algorithms as a decision aid can amplify existing 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 within commuting distance of cities. Kleinberg et al. (2018a) 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 pre-trial 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. Brynjolfsson et al. (2019) provide an example of how the introduction of automated machine translation through Natural Language Processing on an international trade platform significantly increases transactions and thus economic efficiency. With their setting, the authors demonstrate how AI can help overcome barriers to efficiency, in this case, language barriers to international trade. 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. Our study mainly relates to this literature by providing 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.

3. Theoretical Framework

To be able to study the questions at hand, we deliberately choose a controlled, however, abstract setting, where we show how the employment of a (discriminatory) AI system would affect the well-being of actual people. More specifically, we make use of a reduced version of a one-shot sequential prisoners' dilemma that reflects the incentive and informational structure of any one-shot sequential economic exchange that takes place in the absence of enforcement mechanisms, e.g., because they are prohibitively costly or reputation effects are absent (Fehr and Fischbacher 2003, Dufwenberg and Kirchsteiger 2004). Broadly, one may conceive these sequential exchanges as market transactions where reputation and repeated interactions, at least in the short run, cannot serve as an enforcement mechanism (Fehr et al. 1993, Brown et al. 2004). While arguably neglecting several aspects of real-life environments, the abstraction allows a context independent examination efficiency and welfare ramifications of letting AI systems make decisions on behalf of human stakeholders.

Figure 1



Note. The reduced one-shot sequential prisoners' dilemma employed in the current paper.

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). If the trustor chooses to keep the endowment, the game ends and both players earn

their initial endowment, i.e., 10 MU. However, if the trustor chooses to transfer her endowment, the trustee observes the trustor's decision 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).[1]

There are two noteworthy aspects to this structure. First, trustors make their initial strategic decision under uncertainty, not knowing how trustees will respond. Trustees on the other hand possess full information about trustors' choices when deciding. In other words, there exists an information asymmetry. 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. While abstract, this setting allows us to precisely estimate the pure efficiency and welfare effects that the introduction of an intelligent support system may have.

Given the lack of information and the absence of other enforcement mechanisms such as reputation building, 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 and may yield the largest benefits. 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 and thereby the need for (more costly) enforcement mechanisms, bearing the potential for considerable efficiency gains. 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 (e.g. Tversky and Kahneman 1974, Kahneman and Tversky 1977).

In the following, we derive a simple theoretical framework to illustrate the structure of the setting more formally and provide an analytic contemplation of how the deployment of a (discriminatory) AI system may affect broad population-wide outcomes. We will start considering a benchmark where there are no information asymmetries between trustors and trustees, i.e., where there is no need for using a ML algorithm to produce predictions about trustees' behavioral responses. Subsequently, we relax the assumption of perfect information and show formally how and under what conditions ML algorithms help reduce information asymmetries.

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

[1] Note: This reduced form of the sequential prisoners' dilemma resembles the structure of a trust game (Berg et al. 1995)

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 reduced version of a one-shot 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, conditional on the trustor initially cooperating, is equivalently denoted as $A_2 = \{C, D\}$.

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.[2] 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 to defect $a^*_{i,2}(s) = D$ if the trustor chooses to cooperate. r-types in the role of a trustee, on the other hand, behave reciprocally, i.e., $a^*_{i,2}(r) = C$. Notably, this setting implies that the types can only be distinguished if a trustor initially cooperates, mirroring selective labels environments (Lakkaraju et al. 2017). 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 $f(x)$ generating 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 H 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 $f_H(x) = \hat{\theta}$. 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 widely used model by Charness and Rabin (2002). This model allows individuals to have conditional social

[2] 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 $k = 2$.

welfare and altruistic motives by including the material payoff of other individuals as a weighted component into the utility function. The extent of these concerns is reflected in the magnitudes of model parameters. In a recent study, this preference model has been shown to explain empirical observations of sequential prisoners' dilemmas extremely well (see Miettinen et al. 2020), providing us with confidence that the use of the model is justified. 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. Given the aforementioned preferences of r and s-types, we assume that $\sigma(r) \geq \sigma(s)$ and $\rho(r) leq \rho(s)$. The AI system that makes decisions on behalf of trustor is individually calibrated to know the stakeholder's utility function.

In line with standard literature, we model individuals (and the AI system) as being rational according to their utility functions and act as expected utility maximizers. Under this plausible assumption, the chosen strategy a^* ultimately reflects the solution to the optimization problem

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

$p(a_{j,-k}) \in (0,1)$ 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.[3] Given the structure of the game, trustees only decide about cooperation in case the trustor initially decided to cooperate. As a consequence, trustees do not face uncertainty about the trustor's behavior and assign the probability of one to the strategy C.

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 choose not to reciprocate initial cooperation if $\rho(.) \geq \frac{1}{3}$, which is true for s-types, while they choose to reciprocate cooperation if $\rho(.) < \frac{1}{3}$, which is true for r-types.

For simplicity, we assume that neither type of individual gains utility from the opponents' material payoff, in case their own payoff is smaller than the one of the opponent, i.e., $\sigma(s) = \sigma(r) = 0$.

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

This captures the notion that trustors do not receive a positive utility when their initial cooperation is met with defection, i.e., free-riding, because they are exploited by their opponent. Hence, we can rewrite utility function (1) as

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

As a benchmark, we first solve the game assuming each trustor i observes her matched trustee's types θ_j (in addition to the trustee's personal characteristics x_j). Since there are no information asymmetries, we can use simple backward induction to find the subgame-perfect Nash equilibrium. Because in the equilibrium individuals will necessarily use the same type-dependent strategies, we dispense individual indexation. Being able to identify trustees' types and correctly anticipating their conditional response to initial cooperation, trustors, independent of their type θ, strictly prefer to cooperate whenever the trustee matched with them will reciprocate initial cooperation, i.e., is of type r. Otherwise, when trustees are of type s they are strictly better off choosing to initially defect. All proofs can be found in the appendix.

PROPOSITION 1. *Suppose trustors, before making their decision, observe trustees types θ_j so that there do not exist information asymmetries. There exists a unique subgame-perfect Nash equilibria in which*

$$a^*(s) = a^*(r) = \begin{cases} C & \textit{if} \quad \theta_j = r \\ D & \textit{otherwise} \end{cases} \tag{4}$$

describe trustors' equilibrium strategies conditional on their matched trustee's type, and

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

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

$$\omega\left(C, C\right) = \mu_r \tag{6}$$

$$\omega\left(C, D\right) = 0 \tag{7}$$

and the share of outcomes $\omega\left(a_1^*\right)$ where trustors do not cooperate is given by

$$\omega\left(D\right) = 1 - \mu_r \tag{8}$$

Under perfect information, the share of efficient outcomes where both players cooperate is at its maximum. At the same time, there does not occur any free-riding where a trustor's initial cooperation is exploited and not reciprocated by a selfish trustor. This solution constitutes a first-best scenario from the perspective of the trustor, who, in this setting, generally possess a first-mover disadvantage whenever there exists uncertainty about the trustee's response to cooperative behavior. The perfect information case can serve as a benchmark against which we can compare outcomes under uncertainty where ML algorithms can help reduce information asymmetries.

Next, we relax the assumption of perfect information and consider the case where trustors merely observe trustees' personal traits x_j, but not their actual type θ_j. In this scenario, a ML algorithm that produces accurate individual-level predictions about trustees' types can generate additional value by (i) preventing inefficient transaction breakdowns where trustors initially refrain from cooperation, and (ii) helping to avoid free-riding outcomes where trustors, to their personal disadvantage, cooperate while the trustee does not reciprocate.[4]

Given that we introduce information asymmetries, 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, again 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

$$\hat{\mu}_r \geq \frac{1}{2} \tag{9}$$

This result enables us to derive equilibrium predictions for scenarios where trustors use an AI system to make a decision on their behalf.

[4] Note: Given the functional form of our utility function, the free-riding outcome, from a welfare perspective, is strictly preferable over the defection outcome due to the mechanism that initial endowments are always doubled if cooperation occurs, even if it is not reciprocated. Hence, there only a exists a conflict of individual and collective interest from the perspective of the first mover if the trustee does not reciprocate.

The AI system comprises the predictive ML algorithm $f_H(.)$ 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 $f_H(x) = \hat{\theta}$ 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 stakeholder's 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 economic efficiency and social welfare.

PROPOSITION 2. *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*

$$a^*(s) = a^*(r) = \begin{cases} C & \text{if} \quad \hat{\theta} \geq \frac{1}{2} \\ D & \text{otherwise} \end{cases} \tag{10}$$

describe the AI system's equilibrium strategies, and

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

$$a^*(r) = C \tag{12}$$

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

$$\omega\left(C, C\right) = \mu_r \int_{0.5}^{1} q(\hat{\theta}|r) d\hat{\theta} \tag{13}$$

$$\omega\left(C, D\right) = (1 - \mu_r) \int_{0.5}^{1} q(\hat{\theta}|s) d\hat{\theta} \tag{14}$$

and the share of outcomes $\omega\left(a_1^*\right)$ where the trustor does not cooperate is given by

$$\omega\left(D\right) = (1 - \mu_r) \int_{0}^{0.5} q(\hat{\theta}|s) d\hat{\theta} + \mu_r \int_{0}^{0.5} q(\hat{\theta}|r) d\hat{\theta} \tag{15}$$

Proposition 2 depicts that an AI system's capability to produce efficient outcomes depends on it correctly classifying reciprocal trustees as such. 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. Intuitively, as $\int_{0.5}^{1} q(\hat{\theta}|r)d\hat{\theta}$ converges to one, the share of socially most efficient outcomes converges to the perfect information benchmark. However, when the predictive algorithm exhibits a low performance in regards to correctly classifying reciprocal types, it increasingly steers the population away from the first-best outcome by fostering the occurrence of inefficient, welfare minimizing outcomes of trustor defection. In other words, proposition 2 depicts how the recall value of the AI systems predictive ML component determines how useful the system is in terms of facilitating mutual cooperation. On the other hand, an algorithm's performance in correctly identifying actual selfish types determines its ability to prevent free-riding outcomes where trustees exploit trustors' initial cooperation. As $\int_{0}^{0.5} q(\hat{\theta}|s)d\hat{\theta}$ converges to one, the free-riding outcomes cease to occur, i.e., converge toward the perfect information benchmark. These two insights emphasize that a ML algorithm's ability to reduce, or in the best case completely overcome, information asymmetries between trustors and trustees hinges upon its recall performance scores. Notably, the share of efficient and inefficient outcomes does neither depend on accuracy nor precision measures. This constitutes a considerable problem in environments such as the one in our framework, where the observation of an individual's true type depends on the actions the AI system chooses, i.e., a selective labels environment (Lakkaraju et al. 2017), so that the underlying recall value can not be determined. This is because it is inherently difficult to identify the share of false-negative predictions, a central component to the recall performance measure.

Proposition 2 also provides insights into the impact of algorithmic discrimination on economic efficiency. In line with contemplations by Adomavicius and Yang (2019), we refer to algorithmic discrimination as algorithmically generated decisions that are inaccurate and disadvantageous for a specific subgroup in the population. In our setting, we define algorithmic discrimination relative to the benchmark of perfect information. Specifically, a system discriminates if its decisions for a specific subgroup of trustees, to their disadvantage, systematically differ from decisions that would maximize the stakeholder's utility if these trustees' types θ were perfectly observable. This is the case, if the predictive ML algorithm $f_H(.)$ incorrectly produces overly pessimistic predictions that individuals with a specific characteristic $x_k = 1$ are reciprocal (incorrectly low values of $\hat{\theta}$), while predictions for individuals with $x_k = 0$ are accurate, leading to unfairly and inefficiently low cooperation with trustees who have the characteristic $x_k = 1$. The discriminatory outputs may either occur due to explicit programming, or due to incorrectly learned statistical patterns from data H that reflect the discriminatory practices rooted in societies (Berendt and Preibusch 2017). With this notion of discriminatory algorithmic outputs, the recall measure for different subgroups

can serve as a measure for discrimination as it reflects the share of r-types that are actually identified as such. Whenever there exists an economically and statistically significant difference in the recall value of a predictive ML algorithm $f_H(.)$ between trustees based on a specific trait x_k, e.g., male or female, black or white, Christian or Muslim, the system exhibits algorithmic discrimination.

Following this line of argumentation, proposition 2 indicates that the share of efficient outcomes of mutual cooperation decreases, while trustor defection occurs more frequently if a system discriminates against specific subgroups. This is because the share of mutually cooperative (defective) outcomes increases (decreases) with the population-wide recall value, which is a subgroup-weighted average of distinct subgroups' recall scores. Hence, if a system discriminates and therefore has a low recall score for a specific subgroup, the share of efficient outcomes decreases. The magnitude of the drop in efficiency naturally depends on the discriminated subgroup's relative size in the population. Notably, the decrease in the frequency of cooperative outcomes creates losses for trustors and trustees alike, so that it is not only the discriminated group that bears the costs but also the stakeholders of the system.

4. Empirical Investigation: Field Data Collection and Simulation Design

We base our analyses on a rich data set that we collected in an incenitivzed field study over a period of three years between 2016 and 2019. Specifically, at the beginning of each semester, we invited first-semester economics students from a large German University to participate in our study. Most important for the current paper, the field study includes an incentivized one-shot sequential prisoners' dilemma along the lines presented before, allowing us to elicit participants' revealed preferences through their behavior instead of observing mere hypothetical survey responses. We show the exact instructions in the Appendix B. We elicited field study participants' behavior using the strategy method. This is, every participant needed to define an action conditional on the choices of the trustor, providing us the unique opportunity to observe consequences of counterfactual choices that trustors do not actually make.[5] In addition to the incentivized game, the field study 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. We show an overview of all items in the Appendix B in figure 11. We paid participants according to their (and their opponents') choices in the sequential prisoners' dilemma. Specifically, we randomly

[5] Note: In the field study, trustees not only have to decide whether to cooperate or defect if the trustor initially cooperates but also she initially defects. However, for the study at hand, we only use observations where trustees decide to defect if the trustor initially defected. These two types make up for 93% of our usable post-cleansing observations. As a consequence, the game reduces to the form presented in previous sections.

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 monetary unit earned in the game, chosen participants received 1 Euro. On average participants earned 13.16 Euro through their choices.[6] We conducted the study online on *LimeSurvey*.

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. Notably, each observation that we use for the study at hand, 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 as comprehensive empirical testing in regards to feature 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.

Table 1 Items from field study used as features to train the ML algorithm.

	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

Note: We normalized the scale of numeric items 1 to 10 in in the pace of the training processes.

The objective of this paper is to study individual and population-wide efficiency and welfare effects of integrating discriminatory AI systems into human societies that inaccurately produce disadvantageous predictions for a specific group of individuals. To do so, we use our cleaned data as a basis for distinct simulation exercises. Notably, while we simulate the game outcomes, they

[6] Note: Subjects effectively received 1 Euro per MU they earned.

reflect outcomes of an interaction between real people and the AI system. Put differently, we do not model or simulate trustees' choices. Instead we are in a unique position to use actual people's behavior.

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 and Schapire 1997). Our trained algorithms exhibit a high performance on all relevant technical performance measures. Table 3 in the Appendix B shows a performance overview of our algorithms after validation and training on a test set. At this point it is important to emphasize that the type of algorithm we use is not of fundamental importance here. We acknowledge that we could also employ statistical methods such as a logistic regression and that the amount of data we harness to train the model cannot be considered Big Data. The key notion, however, is that there exists a model with reasonably high predictive performance (for a specific subgroup in the population) that produces a forecast which feeds into a larger system of automated decision making. The main insights we intend to generate, namely precisely outlining consequences of algorithmic discrimination and empirically testing implications from our framework, are independent of the type of ML algorithm or statistical model.

Algorithm 1: Sequence of simulation exercises

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 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 the reduced 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, the trained ML component produces predictions about trustees' likelihood to reciprocate cooperation. The decision making component subsequently uses this prediction and individual trustors' previously estimated utility functions[7] to compute whether cooperation or defection yields a larger expected utility. The AI system's decision is the utility-maximizing strategy, which is 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.

We are mainly interested in the ramifications of discriminatory algorithmic outputs, i.e., outputs that are inaccurate for a specific group of individuals (here: women) and thereby lead to unfair and disadvantageous outcomes for these individuals, as compared to individuals not belonging to this group (here: men). To this end, we deliberately manipulate the predictive ML component of AI systems so that it systematically underestimates the probability that a female trustee, relative to a male trustee, will reciprocate cooperation, 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 $p < 0.000$). As a consequence, the AI system will unfairly cooperate less often with women, reducing their potential payoffs. While discriminatory outputs may result from both explicit programming or due to societal patterns incorporated into the data, we introduce the algorithmic discrimination 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.

[7] 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.

The use of the gender attribute should be understood as a representative example of a broad range of characteristics that algorithms may discriminate on. Yet, we choose the gender attribute, together with introducing the discrimination through imbalances in the training set, as an example of algorithmic discrimination to pin down the consequences of discriminatory systems for two reasons. 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 (e.g. Sweeney 2013, Buolamwini and Gebru 2018, O'Neil 2018, Lambrecht and Tucker 2019). Second, male and female participants in our field study 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 algorithmic discrimination in the first place.

Finally, to examine interaction effects between algorithmic discrimination 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.

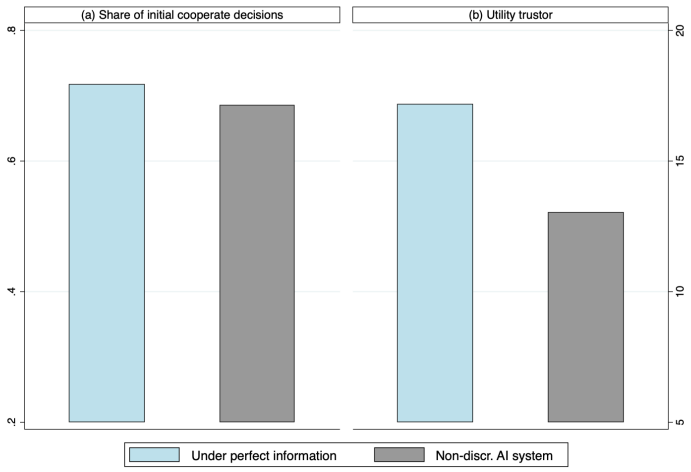
5. Empirical Investigation: Results

We present the results of our simulation exercises in three parts, mirroring our theoretical framework. First, we examine how well a non-discriminatory AI system performs in making trustor decisions, relative to a perfect information benchmark. The perfect information benchmark gives an idea about how well the system bridges information asymmetries between trustors and trustees. These findings serve as a reference point enabling us to outline how results change in case the underlying ML algorithm increasingly discriminates against women. By doing so we show in detail the role algorithmic discrimination plays regarding AI systems' potential to produce efficient outcomes. Finally, we study to what extent continued learning may, over time, enable a strongly discriminatory ML algorithm to recover itself.

5.1. Non-discriminatory AI system

We start our analyses by examining the performance of a non-discriminatory AI system in making decisions on behalf of human trustors. Non-discriminatory refers to the fact that in comparison to subsequent AI systems, we did not intentionally introduce algorithmic discrimination 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 ($p < 0.12$). Notably, even though economically insignificant, the recall value for women is even 6.1 percentage points higher for women compared to men. We initially focus on the system's performance from the perspective of human stakeholders. We will consider two distinct measures. First, we compare the shares of cooperative trustor decisions. Subsequently, we consider differences in average trustor utility across the two scenarios.

Figure 2



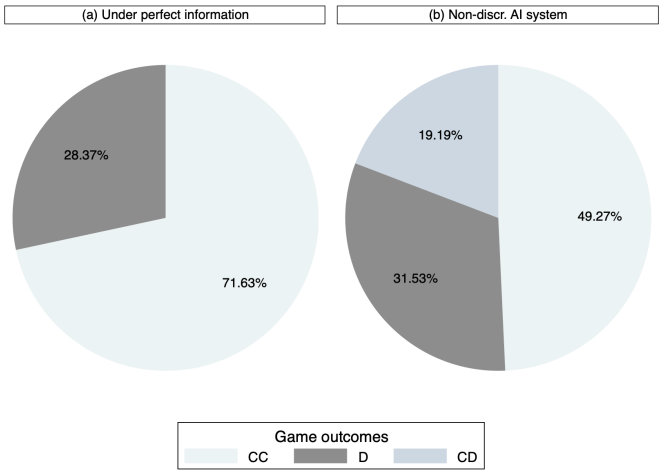
Note. Panel (a) represents the shares of cooperative trustor decisions, while panel (b) shows average trustor utility. Both panels depict results for the benchmark of perfect information and an AI system.

Figure 2 depicts the shares of cooperative trustor decisions (panel (a)) and the average trustor utility the system generates (panel (b)) for both a benchmark of perfect information where trustees' types are observable and the case where an AI system decides on behalf of human stakeholder under imperfect information. We use the perfect information scenario as a benchmark throughout the empirical part of the paper because it indicates how well the AI system performs in overcoming information asymmetries by providing accurate individual-level predictions in an uncertain environment.

Under perfect information, a trustor would cooperate in 71.6% of the cases. In comparison, the AI system cooperates in 68.5% of the cases, which is a difference of 3.1 percentage points. While this difference appears to be rather small, indicating a high performance of the system, a

comparison of the average trustor utilities reveals that there are considerable inefficiencies. On average, the AI system creates a utility of 13 units for human stakeholders. In the first-best scenario the average trustor utility equals 17.16 units. Put differently, the non-discriminatory AI system, from the perspective of their human stakeholders, reaches about 75% (=13/17.16) of the maximum possible utility. Looking at individual choices, we find the AI system to make the same decision that would occur under perfect information in only 58.4% of the games. More specifically, conditional on the trustees' types, the AI system correctly chooses to cooperate (defect) in 80% (29.1%) of the cases. This suggests that the system, while exhibiting similar cooperation rates as under perfect information, frequently chooses to cooperate even though the trustee does not reciprocate, while it also frequently defects even though the trustee would have reciprocated.

Figure 3



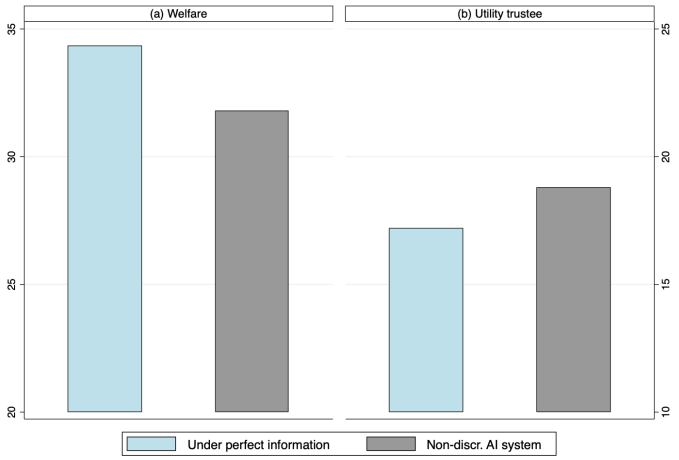
Note. We depict relative frequencies with which different game outcomes occur. Mutually cooperative (CC), trustor defection (D), and free-riding (CD) outcomes. Panel (a) represents bench-mark results; panel (b) represents results for the AI system

Naturally, this pattern of decision making has population-wide consequences for the efficiency of game outcomes. Figure 3 shows the shares of overall game outcomes. Panel (a) depicts outcomes for the benchmark, while panel (b) illustrates the AI system scenario. CC, D, and CD respectively refer to outcomes where trustors and trustees both cooperate, where trustors defect, and where trustors cooperate while trustees defect.

Panel (a) depicts that under perfect information, the socially efficient outcomes of mutual cooperation occurs in about 71.6% of the games, while the share of defective outcomes equals 28.4%. Naturally, there are no free-riding outcomes where initial cooperation is exploited by a trustee under perfect information. Observations in panel (b) depicts that the AI system can only reach

the most efficient outcome in 49.2% of the cases. This is 22.4 percentage points short of the benchmark, implying that in 22.4% of the cases, the system defected even though cooperation would have resulted in the socially and individually most efficient outcome. Put differently, in about one quarter of all games, there is space for a Pareto improvement where trustors and trustees can simultaneously be made better off without making any one of them worse off. On the other hand, in 19.2% of the games, the AI system initially cooperated to the disadvantage of the human stakeholder.

Figure 4



Note. Panel (a) represents average welfare, while panel (b) shows average trustee utility. Both panels depict results for the benchmark of perfect information and an AI system.

Figure 4 depicts the consequences of the AI system's incorrect decisions in terms of average population welfare and average utility trustees gain when interacting with the system. The figure shows that the aforementioned Pareto inefficiencies translate into welfare losses relative to the first best scenario. The AI system reaches an average of population welfare (the sum of trustor and trustee utility) of 31.8 units, which is about 92.7% of the perfect information benchmark. However, one can also see that the relatively small difference to the benchmark is at least in part driven by free-riding outcomes which considerably increase trustees' well-being. Compared to the benchmark case, the average utility of trustees is 1.6 units higher (+9.3%). When excluding the outcomes where social welfare increases due to free-riding, at the expense of the trustor, we find the machine to reach about 87% of the welfare level that occurs under perfect information.

5.2. Discriminatory AI Systems

After we have outlined the performance of an AI system that does not discriminate against women and compared its performance in terms of efficiency and welfare to a benchmark of perfect infor-

mation, we now proceed with the main part of our empirical analyses and outline how previous results change in response to introducing algorithmic discrimination.

We intentionally introduce algorithmic discrimination 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. When an AI system's ML component systematically underestimates the probability that female trustees reciprocate initial cooperation, it will, to females' disadvantage, defect more often when interacting with them. We choose the gender attribute as a basis of discrimination to showcase the consequences on efficiency and welfare.

We create algorithmic discrimination by imbalancing the training set. In imbalanced training 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. At this point, we want to emphasize that the analyses do not aim at discussing the reasons for AI systems to learn discriminatory behavior. Our intention is to outline a clear and precisely quantifiable empirical example of how algorithmic discrimination can negatively impact economic efficiency and welfare, thereby substantiating implications from our theoretical framework.

To see that we successfully introduced algorithmic discrimination, consider table 2 which depicts the average predicted probabilities that women and men cooperate, conditional on the degree of imbalance together with recall scores for men and women. As argued in our theoretical framework, the recall score depicts algorithmic discrimination by revealing that the algorithm makes significantly less capable of correctly identifying female reciprocators as such. 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.

Given that we have successfully introduced algorithmic discrimination in our framework, we now examine to what extent our previous results change when the AI system, to a varying degree, discriminates against women. We start examining how discrimination, measured by the recall scores, influences the AI system's performance from the perspective of the human stakeholder.

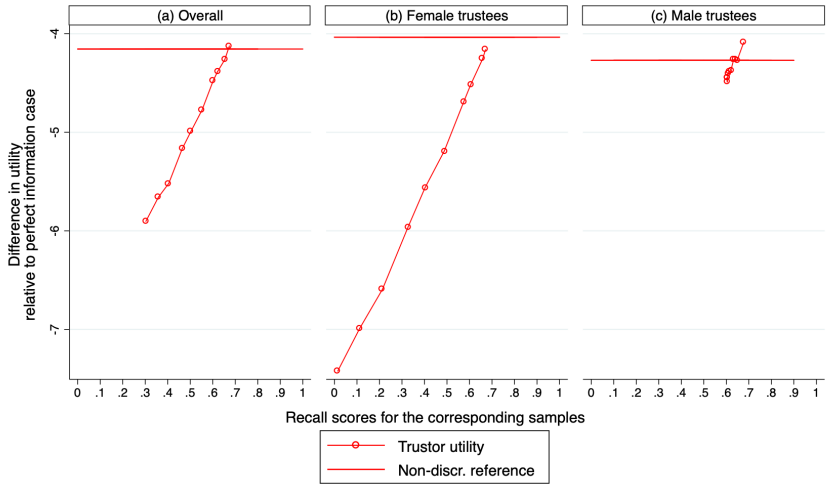
Table 2

	Share of reciprocal examples among female observations in the training set										
	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Recall women	0.02	0.11	0.21	0.33	0.4	0.49	0.57	0.61	0.66	0.67	0.71
Mean pred. women	0.03 (0.107)	0.15 (0.257)	0.25 (0.314)	0.36 (0.344)	0.42 (0.364)	0.5 (0.358)	0.56 (0.357)	0.59 (0.35)	0.64 (0.339)	0.64 (0.33)	0.69 (0.318)
Recall men	0.6	0.62	0.6	0.61	0.61	0.62	0.63	0.64	0.65	0.68	0.66
Mean pred. men	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)

We show the share of reciprocal individuals, mean predicted probabilities together with standard errors, and recall scores of different ML algorithms. The true average share of reciprocal women equals 0.75 (0.435). The true average share of reciprocal men equals 0.69 (0.462). We report standard errors in parentheses.

Subsequently, we outline population-wide efficiency and welfare ramifications. Note that we always consider results relative to the benchmark of perfect information allowing us to see by how much algorithmic discrimination impedes the bridging of information asymmetries between trustors and trustees.

Figure 5



Note. We show results relative to the benchmark of perfect information. From left to right panels show results for (a) the entire sample of games, (b) the subsample of games with female trustees, and (c) the subsample of games with male trustees. Results for a share of 50% of reciprocal observations among female examples in the training data represent the non-discriminatory case from the previous section and are depicted as horizontal line.

Figure 5 depicts the performance of AI systems, relative to the perfect information benchmark, conditional on the degree of discrimination against women. Following our theoretical framework, we measure discrimination by the corresponding (sub)samples' recall scores. Depicted plots show the difference in the average utility of trustors. The horizontal line depicts the reference value of the non-discriminating AI system we explored in the previous section. Negative values on the Y-axes

indicate that the AI system cooperates less (generates less trustor utility) compared to the perfect information case. Positive values would 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 games where trustees are men.

The figure portrays that an AI system's capability to produce outcomes that are utility-maximizing from the perspective of the human stakeholder negatively depends on the degree to which the system discriminates against the subgroup of women in the population. The more inaccurate algorithmic predictions are for women, the less trustor utility the system generates. This observation mirrors the implications of our theoretical framework. Both, the produced trustor utility sharply increase with the recall score, i.e., decrease with the extent of algorithmic discrimination. On average, the trustor utility increases by 0.05 units per percentage point of recall score.

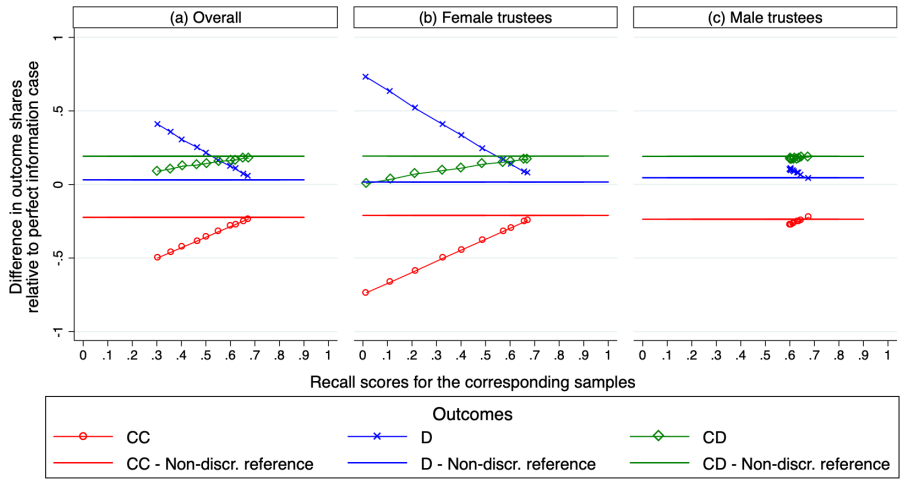
Intuitively, the negative effects of algorithmic discrimination for trustors are driven by instances where the trustee is a woman (see panel (b)) since the predictive performance is low for this group of individuals. When the most discriminatory system (recall value for women is equal to 0.02) makes trustor decisions, the average utility of trustors is about 3.4 units lower compared to the case where the system does not discriminate. This is an economically and statistically significant decrease by more than 25% of trustor utility. Notably, the relation between the recall score, trustor utility is also present when considering the subsample of male trustees, which slightly varies with the imbalance in the female training observations as well.

Overall, these observations provide empirical support for our model implications that we discuss in the theory section of the paper. It is in the interest of trustors that the AI system that decides on their behalf does not unfairly discriminate as their utility decreases with the degree of the extent of discrimination against women.

Result 1 *There exists a strong negative relation between human stakeholders' economic well-being and the extent of discrimination of an AI system that makes decisions on the human's behalf. The more the system discriminates, the worse off is the human stakeholder in terms of utility. An AI system's capability to overcome information asymmetries in favor of the trustor critically depends on the absence of algorithmic discrimination.*

Next, we consider how discrimination against women affects the population as a whole. Figures 6 and 7 respectively illustrate population-wide effects in terms of how the occurrence of game outcomes and population welfare as well as trustee utility diverges from the perfect information baseline when the AI system increasingly discriminates against women. We show results for all games (panel (a)), the subsample of games where trustees are women (panel (b)), and the subsample

Figure 6



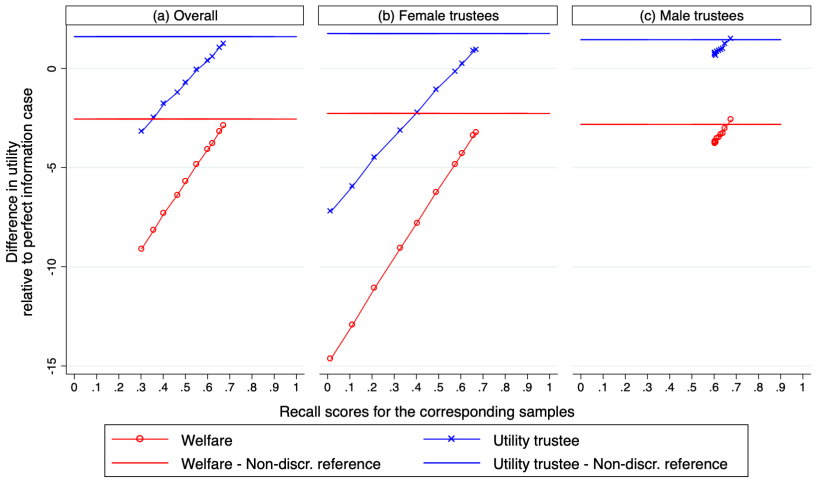
Note. We depict differences in relative frequencies with which specific outcomes occur. We show results relative to the benchmark of perfect information. From left to right panels show results for (a) the entire sample of games, (b) the subsample of games with female trustees, and (c) the subsample of games with male trustees. Results for a share of 50% of reciprocal observations among female examples in the training data represent the non-discriminatory case from the previous section and are depicted as horizontal lines.

of games where trustees are men (panel (c)). Again, horizontal lines depict the reference values of the non-discriminating AI system we explored in the previous section.

Both figures emphasize the detrimental population-wide consequences that the use of discriminatory AI systems may entail. The more a system discriminates, the more it increases (decreases) the occurrence of the socially most efficient (inefficient) outcome (see figure 7). Compared to the non-discriminatory system, the most discriminatory one reaches the mutually cooperative outcome 27.6 percentage points less often (drop from 49.3% to 21.1%), while the occurrence of initial defection is 37.7 percentage points higher (increase from 31.5% to 69.2%). These negative ramifications are largely driven by games where trustees are female. Showcasing the significantly less favorable treatment of women compared to men, in the most discriminatory 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 discriminatory systems' inefficiently low cooperation with female trustees, social welfare decreases substantially. Comparing the system that discriminates most strongly with the one that does not exhibit algorithmic discrimination, welfare subsides by 6.5 units (from 31.8 to 25.3 units) which equals a reduction of 20%. Highlighting the severely unequal treatment in the most discriminatory system, it is the group of female trustees who bear the brunt of the welfare loss since their average utility drops by about 9 units (from 19.2 to 10.2). In comparison, the average trustee utility for men merely drops by 0.6 units.

Figure 7



Note. We depict differences in welfare and trustee utility. We show results relative to the benchmark of perfect information. From left to right panels show results for (a) the entire sample of games, (b) the subsample of games with female trustees, and (c) the subsample of games with male trustees. Results for a share of 50% of reciprocal observations among female examples in the training data represent the non-discriminatory case from the previous section and are depicted as horizontal lines.

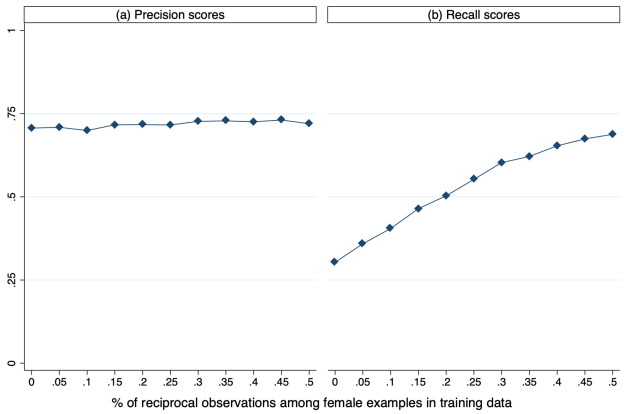
Result 2 *There exists a strong negative relationship between the extent of algorithmic discrimination and economic efficiency as well as population welfare. The discriminated group bears the brunt of the harm. The potential to augment social welfare is inextricably linked to a system's resilience not to inherit discriminatory behavior in the training process.*

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, however, ends the game so that there are no information about the trustee being a reciprocator or not. As shown before, the trustee may be a reciprocator who has falsely been classified as someone who does not respond to cooperation with cooperation without the AI system ever 'finding out'. This selective labels issue (Lakkaraju et al. 2017), reflects the fundamental structure of a multitude of real-life situations in which algorithms 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 et al. 2018a), 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 can compute the real recall performance score. 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



Note. We depict the predictive performance of ML algorithms. We show results conditional on the imbalance in the subset of training examples for women. 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 the imbalance in the subset of training examples for women. Panel (a) depicts the share of utility-maximizing 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 utility-maximizing trustor decisions given that cooperation would have been reciprocated (i.e. the recall score), which is generally not available in real-life scenarios.

The figure depicts an alarming pattern. Independent of the inherent algorithmic discrimination, and thus of the negative efficiency and welfare consequences we outline above, 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 discriminatory and welfare reducing AI system may be incorrectly assessed as performing reasonably well if one bases the evaluation on this metric. In line with our theoretical framework, panel (b) paints a more accurate picture of the AI systems' performance. It highlights that the overall recall score is sensitive to the algorithm's degree of discrimination. The more the system discriminates, the lower is the value of this performance metric. For instance, instead of indicating that the most discriminatory system performs about as well as the non-discriminating one (respective precision scores: 0.71 vs. 0.72), the recall score shows a considerably lower performance for the most discriminatory 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 significance

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 magnified 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. This finding supports arguments by Adomavicius and Yang (2019) that algorithmic discrimination is a complex issue whose resolution requires human involvement for understanding the source of the issue and defining appropriate performance measures.

Result 3 *In an environment of selective labels, the accurate evaluation of algorithmic performance is difficult and prone to be misleading. Standard available performance measures such as the precision score can provide a highly inaccurate picture of AI systems performance.*

5.3. Algorithmic Discrimination and Continued Learning

So far, our empirical results emphasize the problems algorithmic discrimination may produce considering economic efficiency and social welfare. These observations imply that to maximize the potential benefits of AI systems for societies, it is important to further our understanding of how to counteract algorithmic discrimination.

We, therefore, devote the final part of our analyses to studying how algorithmic discrimination endogenously changes if systems continue to learn within an environment where the originally learned discrimination is 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 differ fundamentally. If we interpret this difference as being the result of a change in a non-stationary environment, algorithmic discrimination, 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 and 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 (e.g. Jordan and Mitchell 2015, Elwell and Polikar 2011).

Following this notion, we study the development of algorithmic discrimination over time, when we continuously update the ML component of our AI system using training data 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

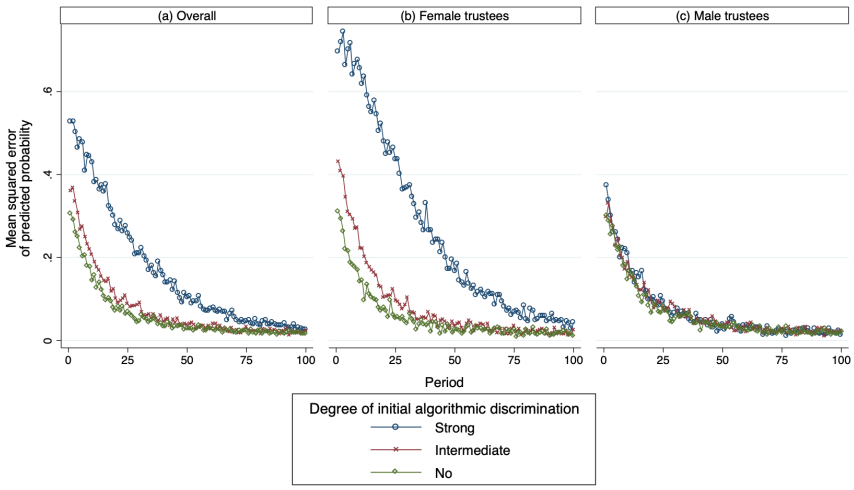
each other (and the AI system) over a certain period. Note that continued learning in our context technically implies that the original training set is increasingly supplemented by a limited number of distinct observations from the population set. As a consequence, the predictive ML algorithm will likely 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 discrimination we initially introduce through distorting the original training data. We consider (i) the non-discriminatory AI system where female examples are balanced with regards to the labels; (ii) an intermediately discriminatory AI system where the share of reciprocal examples among female observations equals 20% ; and (iii) a strongly discriminatory 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 discriminatory patterns are maintained or even reinforced via feedback loops (Cowgill and Tucker 2019).

Since the harmful population-wide consequences of employing discriminatory algorithms stem from systematically incorrect predictions about women's likelihood to reciprocate cooperation, we look at the development of predictions by the ML algorithm over time. More specifically, the development of prediction errors.

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 discrimination 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 discriminatory 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

Figure 9



Note. We depict mean squared errors of predicted probabilities over time. From left to right panels show results for (a) the entire sample of games, (b) the subsample of games with female trustees, and (c) the subsample of games with male trustees.

vs. 0.7) to 0.13. With regards to the intermediately discriminatory 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 discrimination does depend on the extent of the original extent of it. The mean squared error curve in panel (b) for the intermediately discriminatory algorithm is found to be steeper than the one for the strongly discriminatory algorithm. Corresponding curves in panel (c) are virtually identical. This suggests that the algorithm with the intermediate level of discrimination unlearns systematically incorrect predictions for women, in favor of more accurate ones, faster than the algorithm exhibiting the strongest initial discrimination. One explanation, corroborated by our data, is that the less discriminatory 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 discriminatory 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 discrimination. The improved performance and mitigated discrimination naturally translate into positive efficiency and welfare consequences (see figures 12, 13, 14 and 15 in the appendix).

Result 4 *Continued learning can improve ML algorithms' predictive performance over time. Supplementing the training data with affected outcomes and retraining the algorithm can mitigate algorithmic discrimination even in environments of selective labels and if the system's initial level of discrimination is very high.*

6. Discussion and Conclusion

With the paper at hand, we contribute to discussions about the broad consequences of employing discriminatory AI systems. We use both a theoretical framework and an empirical investigation to outline and quantify the potential detrimental efficiency and welfare ramifications of such systems.

Our theoretical and empirical results provide causal evidence that the employment discriminatory AI systems can significantly decrease economic efficiency and social welfare on an individual and a population-wide level. 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 discrimination often originates from historic societal discrimination that is encoded in data, these 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 discrimination is frequently hard to detect so that it may have already 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 and monitoring mechanisms. 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 and calls for human oversight.

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 discriminatory, 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 of Propositions

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 $a_i = C$. If i chooses strategy $a_i = D$, the game ends without j choosing a strategy. There are two types $\theta \in (r, s)$ - reciprocal (r) and selfish (s) - whose preferences are given by

$$U_i(\pi_i, \pi_j, \theta_i) = \begin{cases} (1 - \rho(\theta_i))\pi_i + \rho(\theta_i)\pi_j & \text{if } \pi_i \geq \pi_j \\ \pi_i & \text{otherwise} \end{cases}. \tag{16}$$

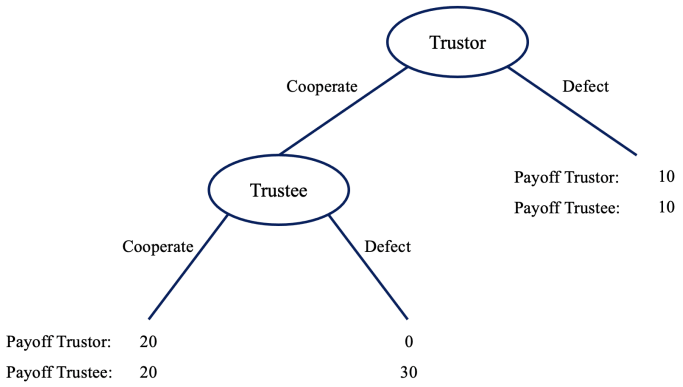
π_i and π_j respectively describe material payoffs earned by the trustor and the trustee. r- and s-types' optimal pure strategies in the role of the trustee, conditional on initial cooperation, are respectively given by $a^*(r) = (C)$ and $a^*(s) = (D)$.

$\hat{\mu}_r$ 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

$$\hat{\mu}_r \cdot U_i(C, C) + (1 - \hat{\mu}_r) \cdot U_i(C, D) \geq U_i(D). \tag{17}$$

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

Figure 10



Note. The reduced one-shot sequential prisoners' dilemma employed in the current paper.

Given the depicted payoff structure, it holds for both types that $U_i(C, C) = 20$, $U_i(C, D) = 0$ and $U_i(D) = 10$. As a consequence, we can rewrite condition (17) as

$$\hat{\mu}_r \cdot 20 + (1 - \hat{\mu}_r) \cdot 0 \geq 10. \tag{18}$$

Proof Proposition 1:

Under perfect information where trustors observe trustees' actual types θ before making a decision, beliefs reduce to unity probabilities conditional on the observed type, i.e., $\hat{\mu}_r = 1$ if a trustee is of type r and $\hat{\mu}_r = 0$ if a trustee is of type s. It follows that whenever a trustor is matched with a trustee of type r initial cooperation is strictly preferred because condition 18 is satisfied ($20 \geq 10$), while defection is preferred if the trustee is of type s because condition 18 is violated ($0 < 10$). As a consequence, the game results in mutual cooperation whenever the trustee is of type r, while the game end with the trustor defecting whenever the trustee is a s-type.

Proof Proposition 2:

Let an AI system comprise the predictive ML algorithm $f_H(.)$ and the codified preferences of the trustor on whose behalf the system decides. $f_H(x) = \hat{\theta} \in (0,1)$ 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

$$\hat{\theta} \cdot 20 + (1 - \hat{\theta}) \cdot 0 \geq 10 \tag{19}$$

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) initial 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.

Table 3

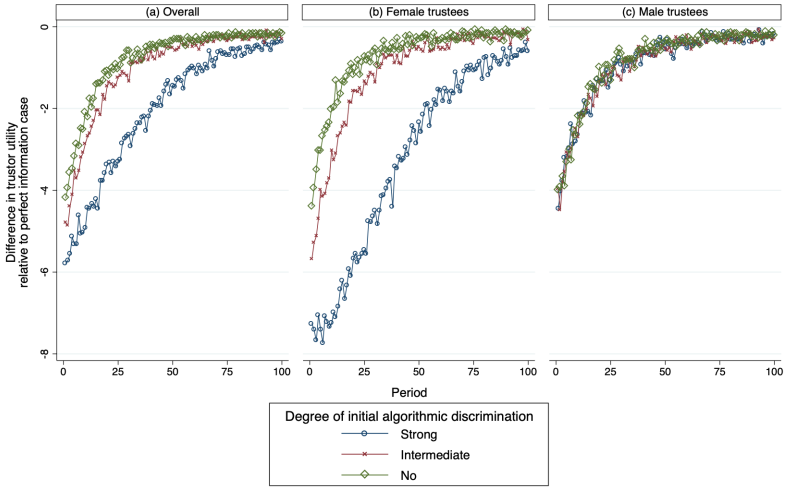
	Share of reciprocal examples among female observations in training set										
Performance measure	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Accuracy	0.84	0.82	0.81	0.76	0.76	0.79	0.75	0.74	0.76	0.76	0.75
Precision (Reciprocal)	0.71	0.71	0.74	0.67	0.71	0.74	0.69	0.72	0.76	0.77	0.77
Precision (Selfish)	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

We how algorithmic performance metrics 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

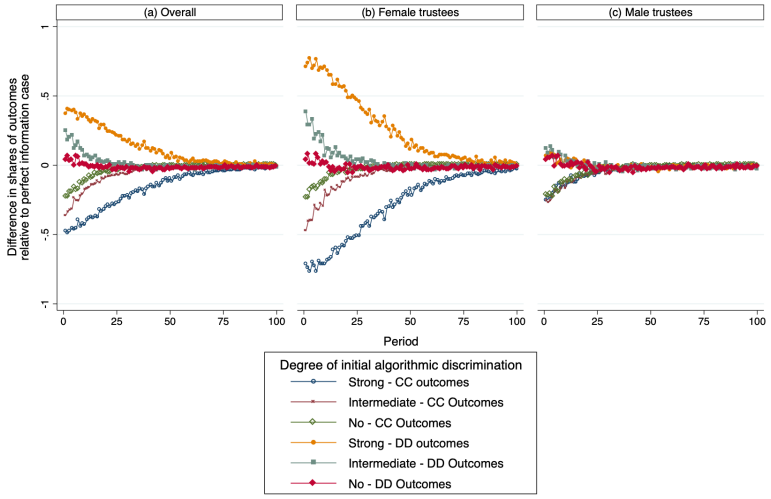
Result: Game outcomes and utilities in sequential prisoners' dilemma games
Cleaning of raw data;
while *counter* $\leq$ *10* **do**
  1. Random partition of cleaned data - 25% population set, 75% training set;
  2. Preparation of training set for training of ML algorithm;
  3. Training, validation, testing 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 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



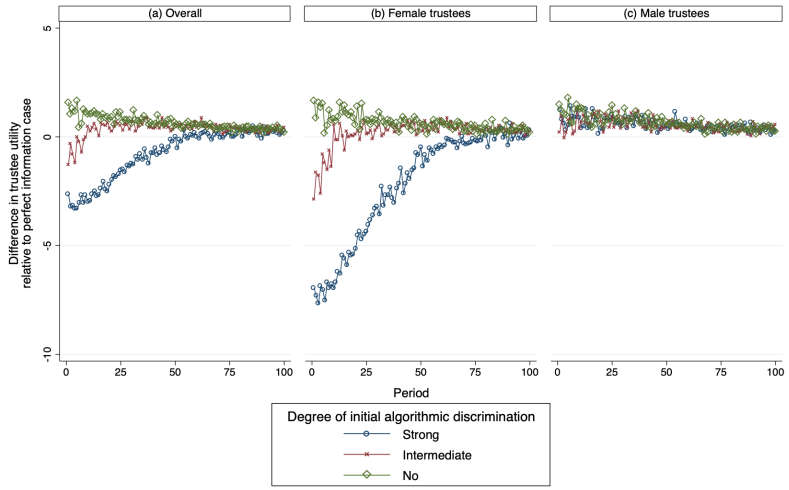
Note. We depict the trustor utility under continued learning. Conditional on the degree of algorithmic discrimination against women. From left to right panels show results for (a) the entire sample of games, (b) the subsample of games with female trustees, and (c) the subsample of games with male trustees.

Figure 13



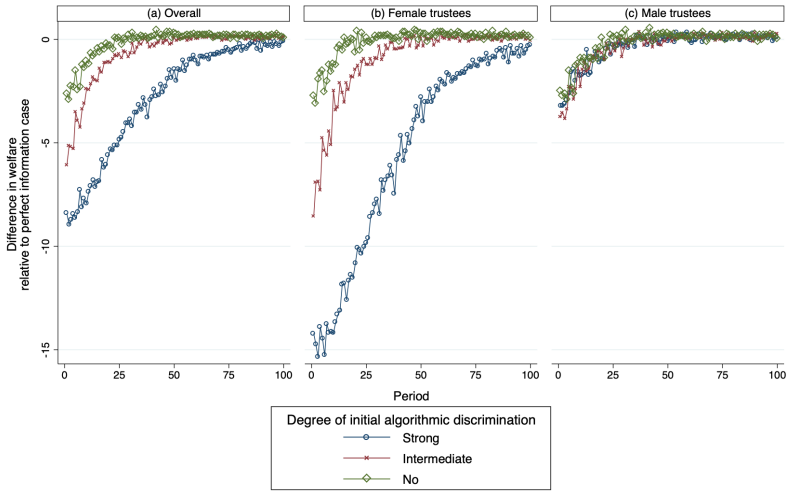
Note. We depict the frequencies with which certain outcomes occur under continued learning. Conditional on the degree of algorithmic discrimination against women. From left to right panels show results for (a) the entire sample of games, (b) the subsample of games with female trustees, and (c) the subsample of games with male trustees.

Figure 14



Note. We depict the trustee utility under continued learning. Conditional on the degree of algorithmic discrimination against women. From left to right panels show results for (a) the entire sample of games, (b) the subsample of games with female trustees, and (c) the subsample of games with male trustees.

Figure 15



Note. We depict population welfare under continued learning. Conditional on the degree of algorithmic discrimination against women. From left to right panels show results for (a) the entire sample of games, (b) the subsample of games with female trustees, and (c) the subsample of games with male trustees.

References

Adewumi AO, Akinyelu AA (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.

Adomavicius G, Yang M (2019) Integrating behavioral, economic, and technical insights to address algorithmic bias: Challenges and opportunities for is research. *Economic, and Technical Insights to Address Algorithmic Bias: Challenges and Opportunities for IS Research (September 3, 2019)* .

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

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. *The Economics of Artificial Intelligence: An Agenda*, 507–547 (University of Chicago Press).

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

Berendt B, Preibusch S (2017) Toward accountable discrimination-aware data mining: the importance of keeping the human in the loop—and under the looking glass. *Big data* 5(2):135–152.

Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games and Economic Behavior* 10(1):122–142.

Bhattacharyya S, Jha S, Tharakunnel K, Westland JC (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.

Brynjolfsson E, Hui X, Liu M (2019) Does machine translation affect international trade? evidence from a large digital platform. *Management Science* 65(12):5449–5460.

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91.

Chaboud AP, 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. Technical report, Technical Report. Working paper.

Cowgill B, Tucker CE (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 SA, 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, Cui C, Corrado G, Thrun S, 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 RE (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 JJ (1979) Sample selection bias as a specification error. *Econometrica* 153–161.

Hendershott T, Jones CM, Menkveld AJ (2011) Does algorithmic trading improve liquidity? *The Journal of Finance* 66(1):1–33.

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

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

Huang CL, Chen MC, Wang CJ (2007) Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications* 33(4):847–856.

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349(6245):255–260.

Kahneman D, Tversky A (1977) Intuitive prediction: Biases and corrective procedures. Technical report, Decisions and Designs Inc Mclean Va.

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

Kleinberg J, Ludwig J, Mullainathan S, Rambachan A (2018b) Algorithmic fairness. *AEA Papers and Proceedings*, volume 108, 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. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 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 TH, 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. URL `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. URL `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 JF, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO, et al. (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.