Humans as Mitigators of Biases in Risk Prediction via Field Studies

Bei Wang
School of Computing and SCI Institute
University of Utah
Salt Lake City, USA
beiwang@sci.utah.edu

Arul Mishra
David Eccles School of Business
University of Utah
Salt Lake City, USA
arul@mishra.us

Himanshu Mishra
David Eccles School of Business
University of Utah
Salt Lake City, USA
himanshu@mishra.us

***Abstract*—Machine learning algorithms have been used for predicting different risks – financial, medical, and legal – and have been argued to perform more efficiently than human experts. However, this exclusive focus on accuracy can be at the cost of the algorithms discriminating against people due to their age, gender, or race, since accuracy could work in opposition to equity. The challenge is that equity and fairness are innately human values that evolve as societies evolve, making it hard to represent them mathematically. Therefore, we propose a framework for including less biased human experts in the algorithm's prediction loop to improve equity and maintain accuracy. In two field studies, one in the legal domain and the other in credit risk, we utilize publicly available datasets to obtain baseline measures of fairness. Subsequently, we obtain human input, which are used to debias the algorithm. Utilizing less biased human experts, as well as providing transparent and explainable predictions, will help increase legal compliance and the trust of various stakeholders in an organization.**

***Index Terms*—Field Studies, Biases, Bias Mitigation, Case Studies and Empirical Investigations, Psychology**

I. Introduction

Risk prediction critically impacts people's everyday lives, which is crucial to many equity and fairness questions in loan approval, criminal justice, and clinical treatment: Will a consumer receive a line of credit? Will a defendant be denied a favorable sentence? Will a patient be given a life-saving treatment? Quantitative models have been used historically for predicting risks and have been argued to perform better than human experts [1]–[4]. The main objective of risk prediction models has been to determine risks accurately, e.g., predicting as accurately as possible whether a borrower will default on payment or a defendant will re-offend. However, this exclusive focus on accuracy has largely ignored the cause of equity, especially with the wider use of machine learning (ML) models for risk prediction that are optimized to achieve high levels of accuracy [5]–[10]. The strength of ML models and the reason for their wider adoption is their ability to identify patterns in the existing data to make accurate risk predictions.

This strength becomes a weakness when existing data include historic human biases, such as race, gender, age, or income-based biases. ML models, unfortunately, represent these biases as relevant rules that humans utilize in their decision-making processes, which leads to a cycle of discrimination, in which models learn biases from past human decisions, and themselves make inequitable predictions. Recently, evidence has emerged that enhancing accuracy comes at the cost of equity [11]–[18]; models end up discriminating against people due to their age, gender, or race, since accuracy could work in opposition to equity [19]–[21].

To address the challenge of balancing accuracy with equity, we propose an interdisciplinary framework that includes input from human domain experts in the algorithmic prediction loop to reduce bias in risk prediction. We use an explainable, rule-based model in risk prediction because it is important that predictions should be transparent, especially in the legal and financial domain in which decisions significantly affect people's every day lives. We depart from approaches that use purely mathematical tools [15], [22], [23] to improve equity and maintain accuracy in risk predictions. We argue that human experts are needed because equity and fairness are innately human conceptions. Moreover, equity reflects evolving societal values – e.g., female participation in the workforce, gay rights, social justice – that societies are continuously redefining. Therefore, input from humans identified as less biased but experts in their domain can help identify and reduce algorithmic bias. Since they are experts in their domains, they would also ensure that the accuracy of the model is not greatly affected while helping reducing any bias in the algorithmic output. In our proposed approach, the input variables are processed to generate decision rules that help determine which variables are considered most impactful by the model. These rules are then provided to unbiased (or, more accurately, less biased) human experts to check whether the rules are fair and accurate. The input from these human experts are then used to update the rules. Unlike a preprocessing step in which the actual input variables are reweighed or relabeled, our approach first analyzes the input to generate decision rules, which are then updated by incorporating human input. Hence, the human expertise is used as an in-processing step in debiasing. If it is identified that an ML model is discriminatory, then technical debiasing procedures are possible at the preprocessing, in-processing, or postprocessing stages. In line with recent calls for human-centered artificial intelligence, which argue that algorithmic methods of debiasing can adversely affect other performance measures and human-in-the-loop debiasing may be a way to address such a technical limitation [24], we

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

propose to debias ML models by using human input.

Analogous to using adversarial learning as an in-processing debiasor [25] in which the discriminator network checks not only for accuracy but also for the fairness of the generator network's output, we use human experts to update the algorithm's output (which, as we describe later, is in the form of decision rules obtained after processing the input variables) for accuracy and equity. In two empirical examinations, one using financial data and the other criminal justice data, we demonstrate that using less biased human experts' input could reduce bias in a rule-based ML model. Importantly, regulations designed to protect consumers from model-based predictions mandate human involvement in mitigating inequity, making our method more aligned with changing regulatory and compliance requirements [15], [22], [23], [26]–[29]. Moreover, regulations ask for transparency in predictions. Since it is difficult to achieve transparency with black-box risk prediction models, we will employ explainable ML models whose decisions can be understood, and hence, are amenable to input from human experts.

In two field studies, we follow the same procedure of obtaining baseline accuracy and equity measures using an explainable prediction model. Then, we obtain input from less biased human experts, e.g., loan officers and legal experts through detailed questionnaire-based profiles created by the algorithm. Finally, the input from less biased human experts were used, as an in-processing procedure, to update model predictions. To preview the conclusion, we find that using human input helps make the ML models more equitable.

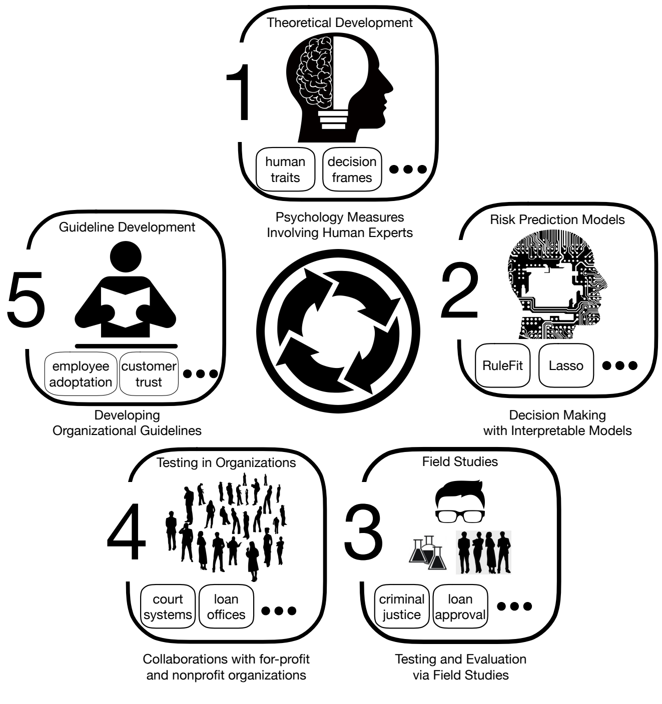

Fig. 1. A framework for identifying less biased human experts and incorporating their input in risk prediction models to balance accuracy and equity.

Our proposed work forms a crucial part of a general framework for identifying less biased human experts and incorporating their input into risk prediction models to balance accuracy and equity, as illustrated in Fig. 1. First, we propose ways to identify less biased human experts who can make a ML model more equitable. Via field studies, less biased human experts will be identified based on quantifiable psychological traits. Second, we incorporate the input of less biased human experts into an explainable, ML model for accurate and equitable decision-making. Third, we evaluate our framework by running multiple field studies using publicly available datasets in the legal and financial domain with evidence of success, which is the main focus of this research. For future work, our study establishes a foundation for collaborating with nonprofit organizations to increase equitable outcomes for unrepresented defendants in cases such as debt collection, small claims, or protective orders; and with for-profit financial services to increase equitable outcomes for credit underwriting and consumer loans. Less biased legal and financial experts can be identified through the method we describe, and their input can be used to update decision rules to provide more equitable outcomes. Our work could also provide organizational guidelines for developing and implementing fair ML models and managing both employee adoption and customer trust within these models.

II. Background and Motivation

Humans as debiasing agents. In using human input to debias a ML model's output, we follow a long tradition of using human judges, also referred to as human-in-the-loop (HITL), as debiasing agents [14], [24], [30]. The HITL literature has demonstrated that humans, lay and experts, can update ML models by providing crucial input for different processing stages as we see in ML models for assisted driving, chatbots, robots, medicine, etc. [31]–[33]. We contribute to research in developing fair ML models by proposing that equitable predictions can be made by identifying less biased human experts using psychological traits since ethics, fairness, and equity are innately human values. Since it is difficult to state that any human is unbiased and biases are implicit, throughout this research, we use the term **less biased human experts** by identifying and treating levels of bias on a continuum. Unlike technical solutions that rely on mathematical tools alone to correct the system, our research contributes by leveraging human input to update an ML model. Therefore, we propose that human expert inclusion is needed to balance accuracy and equity risk because the lack of equity (or bias) is learnt by the model from past biased human decisions. If the humans providing the decisions make errors, then complex statistics is needed to overcome such errors because human error is not driven by random noise but due to some systematic reason [34]. In the case of data that is biased against certain groups, a systematic reason is prejudice. Such systematic reasons introduce bias in risk predictions. Technical solutions depending on mathematical tools alone would be ill-equipped to train ML models to be equitable - an innately human conception that reflects changing societal values. However, assuming expertise to be a sufficient feature could result in including harmful biases in a process that is supposed to debias the system.

Identifying less biased human experts. By the very fact that biases are human, they also vary from one individual to another. Moreover, it has been documented that police use racial profiling in drug arrests [35], and judges and prosecutors apply sentencing guidelines differently to nonwhite defendants [36]. Therefore, we employ psychological measures to identify less biased human experts and use their input to debias the ML model. For instance, two credit underwriters can be comparable in financial expertise but different in terms of the biases they may hold implicitly, e.g., the attitudes they have toward members of a specific group may differ. Since these attitudes are not considered directly relevant for their financial expertise, they are ignored in training the ML models. However, these very attitudes cause biased decisions [37]–[39]. To address such a challenge, less biased human experts are needed to correct the sources of inequity in models. Moreover, findings have demonstrated that purely algorithmic solutions to increase fairness could result in lowered accuracy [24]. Instead, we obtain both accuracy and equity ratings from human experts.

Using personality traits to identify less biased experts. We use innate psychological traits of humans to identify less biased experts. Formally, traits are quantifiable, psychological characteristics, e.g., conscientiousness, impulsivity, frugality, which remain stable over time and reliably predict human motivations and behaviors [40]. Psychological traits can cause one to perceive others differently, given the same information. Personality research has extensively demonstrated that the traits of decision-makers influence whether their decisions are biased or not. In field study 1, we utilize the trait of *openness to experience* (or simply, *openness*) and in field study 2, we use the trait of *need for cognition*, to identify less biased human experts [41], [42].

A. Openness to Experience

First, we hypothesize that **using the input of a high openness human expert would enhance equity in risk predictions**. *Openness* indicates a personality trait of an individual who is willing to make adjustments to existing attitudes and behaviors in response to new ideas, prefers novelty and variety, and is curious, cultured, creative, and less risk-averse [42]. Openness is one of the facets of the Big Five Personality Inventory [43] and has been documented to reduce bias.

Stereotypes have been defined as specific beliefs about a group [44]–[46]. Negative beliefs, referred to as *biases*, have been shown to automatically affect subsequent behavior when certain group information is activated in one's mind [39]. When human experts are high on *openness*, they are willing to accept stereotype-disconfirming evidence, e.g., by associating Blacks with positive descriptions. They also refrain from invoking negative group stereotypes [42]. Such *openness* makes them less likely to use stereotypic associations during their decision-making process, making them less likely to display sexism, racism, ableism, classism, homophobia, transphobia, and xenophobia, resulting in fairer decisions. The 10 scale items associated with *openness*, e.g., "I see myself as someone who is original, comes up with new ideas", or "I am curious about many different things" [43], have been shown to reliably measure the trait. Our field study 1 uses the trait of *openness* to identify less biased human experts.

B. Need For Cognition

Need for cognition (NFC) is a trait that reflects the extent to which individuals are inclined to embrace effortful cognitive activities [47] and has been widely used in psychology [48]. High NFC individuals invest more effort to process the available information compared to low NFC individuals [49]. In field study 2, we measure NFC using participants' responses to 18 scale items proposed by Cacioppo et al. [49], such as whether people agree that the statement "I find satisfaction in deliberating hard and for long hours" describes them well or not. The responses are obtained on a scale from -4 to 4 (-4 being strongly disagree and 4 being strongly agree). Pertinent to bias-detection, low NFC individuals are said to be more likely to use stereotypic information because it is less taxing to use group membership as a decision shortcut [41], [50].

III. Method

We utilize explainable models in our proposed framework, given the requirement by both regulatory bodies and industry to explain decisions made by ML models [51]–[53]. People receiving a prediction expect an explanation of how the decision was reached, especially when there is the potential for harm to protected groups. Some have argued that black box ML models, such as deep neural networks, should be avoided and replaced by models that are inherently interpretable in high-stakes decisions, such as financial risk, healthcare, and criminal justice [54]–[56]. Black box ML models that do not explain their predictions in a human interpretable way [56] have led to severe consequences [57], such as incorrectly denied loans and paroles [54], [55], [58]. Instead of creating methods to explain these black box ML models, a feasible way to obtain transparency is to use ML models that are inherently explainable [56], such as linear regression, logistic regression, decision tree, decision rules, RuleFit, and naive Bayes.

We use RuleFit [59] as a baseline system in our research, since it provides reasons for its predictions in the form of decision rules. RuleFit is an explainable ML model that processes the input variables to generate decision rules. These rules are then provided to less biased human experts to obtain their input as to whether they are fair and accurate (or not). The input from these human experts is used to update the rules. Unlike a preprocessing step in which the actual input variables are reweighed or relabeled, our model first analyzes the input to generate decision rules that are then updated. Hence, **human expertise is used as an in-processing step in debiasing an ML model**. The main idea is to compare rankings of rules based on human-estimated risk and ML model-estimated risk and to use their difference (referred to as *delta ranking*) as a regulatory term in RuleFit; see Fig. 3.

A. Prediction and Decision Rules of RuleFit

RuleFit belongs to a class of ensemble learning algorithms known as prediction rule ensembles. It aims to optimize accuracy as well as explainability by creating ensembles with a small number of simple trees or rules; it also comes with efficient implementations [60]. RuleFit combines the predictions of multiple simple prediction functions to make a final prediction. It consists of a two-step procedure: the first *rule generation* step creates rules from decision trees (i.e., tree ensembles produced by bagged ensembles, random forest, gradient boosting, etc.), and the second *rule fitting* step fits a linear model (such as Lasso) with the original features and the new rules as input [61].

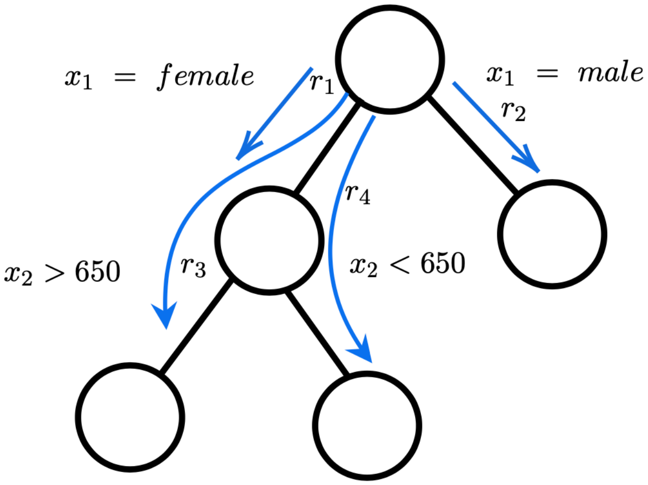

Fig. 2. A tree-based description of how RuleFit works with two variables.

During *rule generation*, RuleFit generates new rules (or features) from decision trees by transforming each path through a tree into a decision rule by combining the split decisions into a rule. Fig. 2 explains this using a simple example of loan approval based on two input features, *gender* (x_1) and *credit score* (x_2). The decision trees in Fig. 2 are trained to predict the outcome variable *loan approval*. The trees show original features as well as their interaction. Four rules (features), r_1, r_2, r_3, and r_4, are generated. r_1 captures whether the applicant is female or not: `IF applicant = female then` r_1 `= 1 ELSE` r_1 `= 0`. r_3 captures the interaction between being female and having a credit score more than 650: `IF applicant = female AND has credit score > 650 THEN` r_3 `= 1 ELSE` r_3 `= 0`. The collection of all such rules derived from all of the trees constitutes a rule ensemble [59].

During *rule fitting*, the rule ensemble is used in prediction models such as L_1-regularized regression or Lasso. This procedure is similar to stacking [62], [63], with the important difference that the members of the ensemble are not learned decision trees or other predictors, but individual rules extracted from trees.

B. Incorporating Traits to Debias Rulefit

To incorporate traits into RuleFit for debiasing, we follow the procedure illustrated in Fig. 3. The two-step procedure of RuleFit provides a set of decision rules that are (Fig. 3a) transparent; one can see what rules RuleFit is mainly basing its predictions on and allows using the input of less biased, human experts to increase equitable risk predictions.

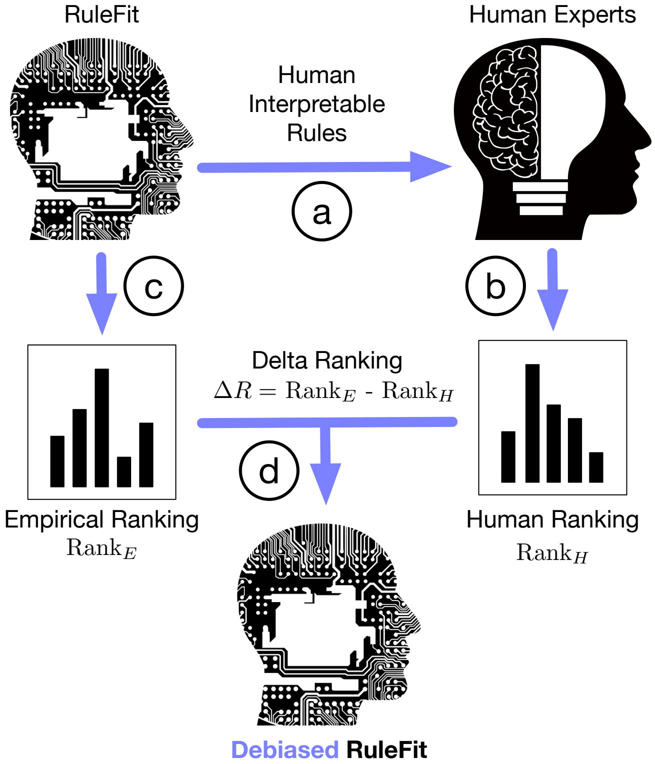

Fig. 3. An illustration of how one debases RuleFit with delta ranking.

For instance, in financial risk prediction as in field study 1, rules learned by RuleFit are converted to borrower profiles and shown to loan officers or credit underwriters. Using the rule: `being male, less than 24 years and with no bank account`, the human expert predicts default risk (i.e., risk of not paying back a loan).

Based on the human expert's response, we capture an expert's assigned predictive importance to this rule, giving rise to a ranking of rules based on human input, referred to as the *human ranking*, denoted as $Rank_H$ (Fig. 3b). RuleFit on its own also assigns importance to each rule, referred to as the *empirical risk*, which gives rise to an empirical risk ranking of the rules (*empirical ranking*), denoted as $Rank_E$ (Fig. 3c). For loan approval, this risk captures default risk of borrowers within the subpopulation defined by the rule. Finally, we define delta ranking [32] $\Delta R = \min(Rank_E - Rank_H)$, which measures the minimum disagreement between human experts and the baseline model (RuleFit) using empirical data. We incorporate bias-corrected human input into RuleFit (Fig. 3d) by identifying rules where $Rank_E$ and $Rank_H$

converge - when human experts and the model have the least disagreement (when their difference is minimized).

Mathematical formulations. Similar to other ensemble learning methods such as bagging predictors [63] and random forests [64], RuleFit [61] takes the form

$$F(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m f_m(\mathbf{x}),$$

where M is the size of the ensemble, $f_m(\mathbf{x})$ is a *base learner* (i.e., a different function of the input variables $\mathbf{x}$ derived from the training data). Ensemble predictions $F(\mathbf{x})$ are taken to be a linear combination of the predictions of each of the base learners, with a_m being the parameters specifying the linear combination. During rule fitting, given a set of base learners, a_m are estimated as $\hat{a}_m$ by a regularized linear regression on the training data $\{\mathbf{x}_i, y_i\}$,

$$\begin{aligned}\{\hat{a}_m\}_0^M = &\arg\min_{\{a_m\}_0^M} \sum_{i=1}^{N} L\left(y_i, a_0 + \sum_{m=1}^{M} a_m f_m(\mathbf{x}_i)\right) \\ &+ \lambda \cdot \sum_{m=1}^{M} |a_m| \qquad (1)\end{aligned}$$

The first term in Eqn. 1 measures the prediction risk on the training data, L is the loss function, and the second term (a regularization term) penalizes large values for the coefficients of the base learners. To incorporate human input to RuleFit, ΔR can be used to replace the regularization term in Eqn. 1 as $\lambda \cdot \sum_{m=1}^{M} \Delta R\,|a_m|$. To obtain *human ranking* ($Rank_H$), let n be the number of human experts, R_m^k the importance assigned by a human expert k to rule m, and P_k an aggregate measure of human bias k. Then an aggregate importance R_m assigned to rule m across n experts will be estimated as in Eqn. 2,

$$R_m = \frac{1}{n} \sum_{k=1}^{n} R_m^k \times P_k. \qquad (2)$$

If values of P_k are inversely related to bias (e.g., *openness to experiences* is low among racially biased individuals), then the above equation assigns less weight to biased humans and provides a bias-corrected value of R_m. When identifying bias through traits such as *openness*, *need for cognition*, or *racism*, P_k will be assessed on a continuous scale.

IV. Field Studies

We next conducted two field studies, each consisting of three stages, as described in Sect. IV-A and Sect. IV-B respectively.

A. Field Study 1: Fair ML Model for Recidivism Prediction

Stage 1: Obtain Baseline Fairness Measures and Generate Human-Interpretable Rules: Stage 1 has two distinct goals: first, to **obtain empirical risk and baseline fairness measures of the ML model without any input from human experts** (Fig. 3c); second, to **generate rules that can be given to human experts to obtain their input** (Fig. 3a). ML models are routinely employed to predict defendant's risk of recommitting a crime. We used a dataset associated with the COMPAS tool, used since 2000, and from a database of 2013-2014 pretrial defendants from Broward County, Florida [9], [16]. Past research has documented significant racial bias in ML predictions made using this dataset [9], [65], [66]. This dataset was randomly divided into 75% training and 25% test sets. To keep the number of rules *human interpretable*, we used features of the defendants as predictors: race (Black vs. White), gender (male vs. female), age (below a median age of 31: yes vs. no), prior convictions (less than a median count of 2: yes vs. no), and charge type (misdemeanor vs. felony). The outcome variable was the *recidivism risk*, that is, a defendant's probability of committing a misdemeanor or felony within 2 years.

We employed a gradient boosting machine (GBM) model to build 500 decision trees with an interaction depth of 4 on the training dataset. We then used RuleFit to transform these trees into a Boolean sparse matrix of rules. Each new rule represents an interaction of original features. For instance, one rule identified by RuleFit as a predictor of recidivism was `Age below 31 = yes AND two or more prior convictions = Yes`. These newly generated rules were used as input features in a Lasso regression model to select the most important ones for predicting recidivism risk. Our analysis yielded the most important 8 rules, and their importance was captured by their *empirical risk*, i.e., the recidivism rate of defendants within the subpopulation defined by a rule. Predicting empirical risk using the 8 selected rules as predictors on the test dataset achieved an AUC-ROC (Area Under The Curve and Receiver Operating Characteristics) of 0.69, which was close to the 0.7 AUC-ROC value, the model performance measure, achieved in previous work using all the features of the COMPAS dataset [16].

We obtained *baseline fairness measures* without the input of human experts and calculated Statistical Parity Difference (SPD) and Disparate Impact (DI) [15], [67]–[69]. According to SPD, a prediction system is unbiased if it classifies the same proportion of individuals from privileged (e.g., Whites) and unprivileged (e.g., Blacks) groups as positive (e.g., not going to recidivate). If there is a disparity between the groups, then SPD is less than zero, which suggests inequity in predictions for the unprivileged group. DI compares the proportion of individuals in the unprivileged group who received a favorable prediction to the proportion of individuals in the privileged group who received a favorable prediction. As per DI, the positive prediction for any unprivileged group should be at least 80% of the rate for the privileged group. SPD was -0.37 and DI was 56%, indicating that the baseline AI system held significant bias against Black defendants. Among many equity measures, our use of SPD and DI is based on consultations with our for-profit and nonprofit collaborators.

Stage 2: Using Field Studies to Identify and Obtain input From Less Biased Human Experts: In Stage 2, we **conducted a field study to identify and obtain recidivism risk predictions from less biased human experts** (IRB 00134035)

(Fig. 3b). The 8 rules derived from Stage 1 were used to develop *defendant profiles*. Participants with prior experience serving as jurors on a criminal trial were recruited by the online panel provider Cint using its advanced prescreening procedure for $4.20 per participant. Only the final sample of 100 participants who had passed each of the attention checks took part in the field study. The median age of participants was 40.5 years, with 47% female; 68% White, 11% African American, and the rest Asian, Pacific islander and Hispanics. Their education levels ranged from high school diplomas to doctoral degrees with a majority (34%) of them having a 4-year college degree.

Participants were first provided with the definitions of various criminal justice terms (e.g., felony, defendant, conviction, misdemeanor, risk), and saw the defendant profiles only after passing the quiz on these terms. Each participant saw 8 defendant profiles, randomly presented, that captured each of the 8 rules identified in Stage 1. For example, one rule `Gender = Male AND Age below 31 = yes AND two or more prior convictions = Yes AND type of crime charged = felony` was converted as the following defendant profile: "The defendant is male and younger than 31 years. He has been charged with felony. He has been convicted of two or more prior crimes." After seeing each profile, participants ranked recidivism risk: "What is the risk of this defendant committing another crime within 2 years?" on a 1 (No Risk) to 5 (Extremely High Risk) scale.

Finally, we measured the participants' personality traits with respect to *openness* by asking them to rate 10 scale items associated with openness [43] on a 1 (strongly disagree) to 5 (strongly agree) scale. Higher average response values indicated more *openness*, and hence, less racial bias. As described in Eqn. 2, the openness value of each participant was multiplied with their respective recidivism risk predictions for each defendant profile. For each profile, the weighted recidivism risk assessment across participants was averaged providing an estimate of R_m for rule m as per Eqn. 2 where lower weight is given to racially biased human experts. Since each profile was based on one of the 8 rules, it allowed us to rank which rules participants have found to be most (or least) predictive of recidivism risk. This is the value of $Rank_H$ described in Fig. 3.

Stage 3: Incorporating input of Less Biased Human Experts to Debias a Risk Prediction Model: Stages 1 and 2 yield two rankings of rules in terms of their influence on the outcome variable: risks assessments with and without human expert input. In Stage 3, we **combine empirical and human risk assessments to make risk prediction model equitable** (Fig. 3d). Following Fig. 3, we estimated the delta ranking ΔR and found the closest convergence, i.e., the least disagreement between human risk ranking and empirical risk ranking, for the two rules: `Age below 31 = yes AND two or more prior convictions = Yes`, and `Race = White AND two or more prior convictions = No`. Human experts ranked the first rule as 4^{th} and the second rule as 5^{th}. Empirical risk estimated by RuleFit ranked them 6^{th} and 3^{rd}, respectively. Using these two rules as input features, we predicted recidivism on the test data. We obtained an AUC-ROC of 0.64. To assess fairness with human experts input, we calculated SPD, which was -0.16 and DI was 81.18%. These values show improvement over the baseline SPD $= -0.37$ and DI $= 56\%$, without human expert input, thus providing support for our framework.

B. Field Study 2: Fair ML Model for Credit Risk Prediction

We conducted a second field study, using the same procedure as the first field study, but in the context of credit risk prediction, to test whether age was used to discriminate among subpopulations. We used *need for cognition* to identify less biased experts.

Stage 1: The credit risk dataset [70] was randomly divided into 75% training and 25% test sets. The following features of the borrowers were used as predictors: gender (male vs. female) and age (above a median age of 24: yes vs. no). The variables with yes vs. no values used were: home ownership, foreign worker, delay in loan payment, guarantor, checking account, employed, real estate ownership, more than 3 people dependent on the borrower. The outcome variable was the *credit worthiness* (yes vs. no). The same procedure (GBM model and RuleFit and Lasso) was used to yield the most important 4 rules. The importance of the rules was captured by their *empirical risk*, i.e., the default rate of applicants within the subpopulation defined by a rule. Predicting empirical risk using all the 4 rules as input features on the test dataset achieved an AUC-ROC of 0.69. Statistical Parity Difference (SPD) was -0.60 and Disparate Impact (DI) was 0%, indicating that the baseline model held significant bias against young applicants.

Stage 2: Our field study used the 4 rules from Stage 1 as *borrower profiles*. Loan underwriters were recruited by Cint using its advanced prescreening procedure for $13.5 per participant. The final sample consisted of 48 loan underwriters. The median age of participants was 36 years, with 72.9% male, 77% White, 10% African American, and the rest were Asians, Pacific Islanders, and Hispanics. A majority (77%) had a 4-year college degree or higher. Each participant was shown profile of borrowers and rated borrower's credit risk using a scale from 1 (No Risk) to 5 (Extremely High Risk). We measured each participant's *need for cognition* on 18 items [43] using a 1 to 9 scale. Higher response values indicated higher *need for cognition*, and hence, less bias toward any group. As per Eqn. 2, the need for cognition value of each participant was multiplied with their respective credit risk predictions for each of the 4 borrower profiles. For each profile, the weighted credit risk assessment across participants was averaged, which provided an estimate of R_m for rule m as per Eqn. 2 where lower weight is given to biased underwriters.

Stage 3: Two rules, using rankings from Stages 1 and 2, indicating the least disagreement between human risk ranking and empirical risk ranking were used as input features to predict credit risk on the test data and resulted in an AUC-ROC of 0.67, a slight drop from AUC$= 0.69$ obtained in Stage 1. Updating with HITL input yielded SPD of -0.12 and DI of 80%, indicating improvement over the baseline SPD $= -0.60$ and DI $= 0\%$, providing further support for our framework.

V. Conclusions and Implications

Our findings across the two field studies using *less biased* human experts to debias the ML algorithm demonstrate our contribution, that is, developing fairer algorithms in order to derive clear policy guidelines. Our field studies demonstrate that using a glass-box model such as Rule-Fit allows users to obtain insights into the model predictions, thus supporting model refinement when the predications are unfair. In particular, we obtain human input (e.g., from credit underwriters or legal experts) to debias the predictions of an ML algorithm. We do not assume that all humans would be equally unbiased and utilize psychological measures to identify less biased human experts and obtain input from them. In summary, we present a framework that utilizes less biased humans in the algorithmic loop to make algorithmic predictions fairer. Our approach is a deviation from the socioculturally-agnostic mathematical abstractions that dominate the literature, by directly integrating human feedback into the model-building process instead of attempting to quantify fairness outright. Furthermore, our rule-based approach may be useful in catching non-robust/spurious rules generally, not just inequitable ones.

However, we acknowledge a few limitations in our framework. We focus on Statistical Parity Difference and Disparate Impact as our fairness metrics and propose that other metrics (such as Equal Opportunity Difference or Average Odds Difference) should also be calculated to evaluate the algorithm's outcomes. Based on the context and availability of data, discussion on which metrics might be most suitable at identifying bias should be a consideration. Moreover, many of these metrics require the data to include protected attributes (such as gender, race, or age) for their calculation. However, many organizations remove these variables from their datasets under the assumption that removing protected variables would remove any potential bias. However, it has been shown that algorithms have the ability to infer these protected attributes from proxy variables and such a practice may result in unfairness-by-unawareness. Finally, using psychological testing to identify less-stereotypically-prone individuals may be one part of the identification process. Other tests can also be conducted or measures put in place so that individuals do not fall prey to "system 1" biases (e.g., conjunction fallacy, availability heuristics, etc.) that occur without one's awareness and can result in bias.

Our findings have several implications. Lack of equity in prediction models is not just a modeling issue; it also has important implications for society and organizations such as business ethics [71] and corporate social responsibility. A coalition of 200 CEOs from the world's leading organizations have argued that companies should be actively monitoring the social impact of emerging AI systems to reduce the influence of systemic social inequalities [72]. Attempts to generate equitable risk predictions face several challenges [11] because even now there is a lack of awareness in organizations about equity issues related to prediction models. Employees have limited time and resources to be devoted to developing their own solutions to creating equitable ML models. Most importantly, this lack of awareness results in employees not being rewarded by the organization for working on equity issues.

Therefore, the findings of our research help us address the important policy implication of how algorithms can be trained via human input to become less biased. It is an interesting circle in which algorithms inadvertently learn biases from historical social biases, i.e., past human biases, and we propose a way of using less biased humans for debiasing the algorithms. Our method is more adaptive because the definition of fairness, an innately human conception, can change with time. Hence, using humans in the algorithmic decision loop can improve the fairness of ML algorithms according to the most recent social definition of fairness.

Acknowledgment

The authors would like to thank the Utah Board of Higher Education's Deep Technology Initiative for supporting the project "Bringing Fairness in AI to the Forefront of Education".

References

[1] R. M. Dawes, D. Faust, and P. E. Meehl, "Statistical prediction versus clinical prediction: Improving what works," *A handbook for data analysis in the behavioral sciences: Methodological issues*, pp. 351–367, 1993.

[2] ——, "Clinical versus actuarial judgment," *Science*, vol. 243, no. 4899, pp. 1668–1674, 1989.

[3] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano *et al.*, "The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[4] V. L. Quinsey, G. T. Harris, M. E. Rice, and C. A. Cormier, *Violent offenders: Appraising and managing risk.* American Psychological Association, 2006.

[5] R. I. Ogie, J. C. Rho, and R. J. Clarke, "Artificial intelligence in disaster risk communication: A systematic literature review," in *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM).* IEEE, 2018, pp. 1–8.

[6] B. T. Pham, A. Jaafari, M. Avand, N. Al-Ansari, T. Dinh Du, H. P. H. Yen, T. V. Phong, D. H. Nguyen, H. V. Le, D. Mafi-Gholami *et al.*, "Performance evaluation of machine learning methods for forest fire modeling and prediction," *Symmetry*, vol. 12, no. 6, p. 1022, 2020.

[7] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PloS one*, vol. 12, no. 4, p. e0174944, 2017.

[8] X. Zhang and S. Mahadevan, "Ensemble machine learning models for aviation incident risk prediction," *Decision Support Systems*, vol. 116, pp. 48–63, 2019.

[9] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, vol. 23, p. 2016, 2016.

[10] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[11] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–16.

[12] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" in *Advances in Neural Information Processing Systems*, 2018, pp. 3539–3550.

[13] N. Kallus and A. Zhou, "Residual unfairness in fair machine learning from prejudiced data," *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, pp. 2439–2448, 2018.

[14] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in Neural Information Processing Systems*, 2016, pp. 4349–4357.

[15] D. Pessach and E. Shmueli, "Algorithmic fairness," *arXiv preprint arXiv:2001.09784*, 2020.

[16] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, no. 1, p. eaao5580, 2018.

[17] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the compas recidivism algorithm," *ProPublica*, 2016.

[18] R. Fu, Y. Huang, and P. V. Singh, "AI and algorithmic bias: Source, detection, mitigation and implications," *Detection, Mitigation and Implications (July 26, 2020)*, 2020.

[19] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[20] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably unequal? the effects of machine learning on credit markets," *The Effects of Machine Learning on Credit Markets (October 1, 2020)*, 2020.

[21] A. L. Hoffmann, "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse," *Information, Communication & Society*, vol. 22, no. 7, pp. 900–915, 2019.

[22] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[23] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 35–50.

[24] S. R. Pfohl, A. Foryciarz, and N. H. Shah, "An empirical characterization of fair machine learning for clinical risk prediction," *Journal of biomedical informatics*, vol. 113, p. 103621, 2021.

[25] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

[26] G. D. P. Regulation, "General data protection regulation (GDPR)," *Intersoft Consulting, Accessed in October 24*, vol. 1, 2018.

[27] "Algorithmic Accountability Act of 2019," https://www.congress.gov/bill/116th-congress/senate-bill/1108, accessed: 2020-12-10.

[28] A. HLEG, "High-level expert group on artificial intelligence: Ethics guidelines for trustworthy AI," *European Commission, 09.04*, 2019.

[29] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[30] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in NLP," *arXiv preprint arXiv:2005.14050*, 2020.

[31] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[32] E. D. Gennatas, J. H. Friedman, L. H. Ungar, R. Pirracchio, E. Eaton, L. G. Reichmann, Y. Interian, J. M. Luna, C. B. Simone, A. Auerbach *et al.*, "Expert-augmented machine learning," *Proceedings of the National Academy of Sciences*, vol. 117, no. 9, pp. 4571–4577, 2020.

[33] D. Xin, L. Ma, J. Liu, S. Macke, S. Song, and A. Parameswaran, "Accelerating human-in-the-loop machine learning: Challenges and opportunities," in *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, 2018, pp. 1–4.

[34] R. Munro, *Human-in-the-loop machine learning*. Manning Publications, 2019.

[35] C. R. Epp, S. Maynard-Moody, and D. P. Haider-Markel, *Pulled over: How police stops define race and citizenship*. University of Chicago Press, 2014.

[36] R. S. Frase, "What explains persistent racial disproportionality in minnesota's prison and jail populations?" *Crime and Justice*, vol. 38, no. 1, pp. 201–280, 2009.

[37] A. G. Greenwald and L. H. Krieger, "Implicit bias: Scientific foundations," *California Law Review*, vol. 94, no. 4, pp. 945–967, 2006.

[38] A. G. Greenwald, B. A. Nosek, and M. R. Banaji, "Understanding and using the implicit association test: I. an improved scoring algorithm." *Journal of personality and social psychology*, vol. 85, no. 2, p. 197, 2003.

[39] P. G. Devine, "Stereotypes and prejudice: Their automatic and controlled components." *Journal of personality and social psychology*, vol. 56, no. 1, p. 5, 1989.

[40] G. Matthews, I. J. Deary, and M. C. Whiteman, *Personality traits*, 2nd ed. Cambridge University Press, 2003.

[41] M. T. Crawford and J. J. Skowronski, "When motivated thought leads to heightened bias: High need for cognition can enhance the impact of stereotypes on memory," *Personality and Social Psychology Bulletin*, vol. 24, no. 10, pp. 1075–1088, 1998.

[42] F. J. Flynn, "Having an open mind: the impact of openness to experience on interracial attitudes and impression formation," *Journal of personality and social psychology*, vol. 88, no. 5, p. 816, 2005.

[43] R. R. McCrae and P. T. Costa Jr, "Personality trait structure as a human universal," *American psychologist*, vol. 52, no. 5, p. 509, 1997.

[44] S. T. Fiske, A. J. Cuddy, P. Glick, and J. Xu, "A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition." *Journal of personality and social psychology*, vol. 82, no. 6, p. 878, 2002.

[45] S. Bhatia, "The semantic representation of prejudice and stereotypes," *Cognition*, vol. 164, pp. 46–60, 2017.

[46] A. R. Todd, A. J. Simpson, K. C. Thiem, and R. Neel, "The generalization of implicit racial bias to young black boys: Automatic stereotyping or automatic prejudice?" *Social cognition*, vol. 34, no. 4, pp. 306–323, 2016.

[47] J. T. Cacioppo and R. E. Petty, "The need for cognition," *Journal of Personality and Social Psychology*, vol. 42, no. 1, pp. 116–131, 1982.

[48] R. E. Petty, P. Briñol, C. Loersch, and M. J. McCaslin, "The need for cognition," in *Hanbook of individual differences in social behavior*. The Guilford Press, 2009, pp. 318–329.

[49] J. T. Cacioppo, R. E. Petty, and C. Feng Kao, "The efficient assessment of need for cognition," *Journal of personality assessment*, vol. 48, no. 3, pp. 306–307, 1984.

[50] C. Stern, T. V. West, J. T. Jost, and N. O. Rule, "The politics of gaydar: Ideological differences in the use of gendered cues in categorizing sexual orientation." *Journal of Personality and Social Psychology*, vol. 104, no. 3, p. 520, 2013.

[51] Executive Office of the President, "Big data: A report on algorithmic systems, opportunity, and civil rights," 2016.

[52] The European Parliament and the Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016," *Official Journal of the European Union*, 2016.

[53] B. Goodman and S. Flaxman, "EU regulations on algorithmic decision-making and a "right to explanation"," *ICML workshop on human interpretability in machine learning (WHI)*, 2016.

[54] R. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer-lending discrimination in the fintech era," National Bureau of Economic Research, Tech. Rep., 2019.

[55] W. Dobbie, A. Liberman, D. Paravisini, and V. Pathania, "Measuring bias in consumer lending," National Bureau of Economic Research, Tech. Rep., 2018.

[56] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intellignece*, vol. 1, pp. 206–215, 2019.

[57] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big Data*, vol. 10, no. 5, 2016.

[58] R. Wexler, "When a computer program keeps you in jail: how computers are harming criminal justice," *New York Times*, June 2017.

[59] J. H. Friedman, B. E. Popescu *et al.*, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2008.

[60] M. Fokkema, "Fitting prediction rule ensembles with R package pre," *arXiv preprint arXiv:1707.07149*, 2020.

[61] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/, 2020.

[62] D. H. Wolpert, “Stacked generalization,” *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[63] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[64] ——, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[65] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[66] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, “Algorithmic decision making and the cost of fairness,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 797–806.

[67] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 329–338.

[68] Y. Awwad, R. Fletcher, D. Frey, A. Gandhi, M. Najafian, and M. Teodorescu, “Exploring fairness in machine learning for international development,” CITE MIT D-Lab, Tech. Rep., 2020.

[69] S. Corbett-Davies and S. Goel, “The measure and mismeasure of fairness: A critical review of fair machine learning,” *arXiv preprint arXiv:1808.00023*, 2018.

[70] H. Hofmann, “Statlog (German Credit Data) Data Set,” *UCI Repository of Machine Learning Databases*, 1994.

[71] S. Tiell, “Create an ethics committee to keep your AI initiative in check,” *Harvard Business Review*, 2019.

[72] M. Nkonde, “Is AI bias a corporate social responsibility issue?” *Harvard Business Review*, 2019.