Human-Machine Collaborative Decision-making in Organizations:
Examining the Impact of Algorithm Prediction Bias on Decision Bias and Perceived Fairness

Anh Luong[a], Nanda Kumar[b], Karl R. Lang[b]

[a] Warwick Business School, University of Warwick
[b] Zicklin School of Business, Baruch College, City University of New York

Working Paper
September 2021

1. Introduction

Recently, algorithmic bias has attracted explosive attention in both industry and academia. Extant research has mostly approached examining algorithmic bias from a technical perspective, with many studies attempting to design fair algorithmic models (Khademi et al., 2019; Xu et al., 2020), collect better quality data (Lu et al., 2019), and operationalize different types of fairness for AI models (Kasy & Abebe, 2020). Few studies have empirically investigated the decision bias of human-machine collaborative decision-making, wherein human experts have the final say after working with the algorithms.

The importance of investigating this topic is twofold. First, human-machine collaboration serves as an important decision-making paradigm in many organizations today (Autor et al., 2019; Daugherty & Wilson, 2018). Particularly, human experts are increasingly employed to work alongside AI in many complex decision-making contexts where the current industry standard of AI predictive power is far from a level that can warrant full automation. Further, societal demands for accountability, regulations, and ethical values require humans to stay in the loop with the machines even in decision contexts with significantly higher AI predictive power. Second, besides machines, research has shown that humans make many biased decisions themselves (Bertrand & Mullainathan, 2004; Tversky & Kahneman, 1974).

Furthermore, studies have also empirically shown that perceptions of decision fairness differ among people, depending on the decision context, personal ethical values, and the roles people take in each context (e.g., whether they are directly affected by the decisions, are the decision-makers, or are the designers of the decision-making model), how they are personally rewarded/penalized based on the decisions they make and so on. It is therefore crucial to move beyond managing the bias of the algorithm component alone to also consider the perceptions and biases of the human component (Adomavicius & Yang, 2019) and, importantly, to examine the bias of the human-machine collaborative decision-making entity as a whole.

The few studies in extant literature that investigated decision bias in human-machine collaboration (e.g. Rhue 2019, Vaccaro, 2019) have not looked at whether different levels of AI bias have varying effects on the decision bias of the human-machine teams. This is important to examine because of two reasons. First, the degree of AI bias often fluctuates, making it crucial to understand the varying effects of different levels of AI bias on the decision bias of human-machine teams. Second, it has not been empirically studied whether technical approaches of bias mitigation from the algorithm side actually mitigate bias in algorithmic decision-making wherein human experts still play a major role. Furthermore, most of these studies have also not demonstrated whether different levels of AI bias have varying impacts on human decision-makers' perceptions of fairness. Understanding this can potentially further explicate decision-makers' behaviors when working with AI in the examination of human-machine teams' decision bias. It can also help organizations better motivate their employees and sustain employees' positive morale for collaborating with the machines. Lastly, most of these studies have not considered the possible exposure effect over time resulting from human decision-makers' repeated interactions with the AI. This is worth investigating because studies in the firm progress function literature (Dutton & Thomas 1984, Lundberg 1961) show that many organizational processes significantly improve in efficacy over time, simply through the act of being carried out by employees repeatedly (exposure).

Thus, we wish to examine the following research questions, specifically in the context of human-machine collaboration for complex organizational decision-making: **RQ1:** What is the effect of AI prediction bias (error rate imbalance[1] between groups) on firm profit and human-machine teams' decision bias? **RQ2:** What is the effect of AI prediction bias (error rate imbalance between groups) on human decision-makers' perceied fairness**? RQ3:** What is the effect of exposure (number of decision periods where decision-makers work with the AI) on organizational profit and human-machine teams' decision bias)?

[1] i.e., inequality

To address these three questions, we conduct a controlled economic experiment with a repeated-round design. We assign participants with a task that models a complex organizational decision-making process wherein human decision-makers (DMs) work with an AI repeatedly over 10 decision periods to evaluate consumer loan applications. We use loan data from a large-scale, historic dataset from *Lending Club* and manipulate the AI predictions to create two experimental conditions: (1) Prediction Bias, where DMs work with AI predictions that discriminate against one group of loan applicants and favor another, and (2) No Bias, where DMs work with AI predictions that treat the two loan applicant groups equally.

We find that human DMs through increasing exposure with the AI learn to adapt to a biased algorithm, implicitly detect the bias in the AI, adjust their behavior, improve significantly their performance, and most importantly, outperform the biased AI working alone in terms of reducing decision bias and increasing organizational profit.

2. Related Studies

We review here the experimental studies examining decision bias in the realm of human-machine collaborative decision-making. Rhue (2019) showed that biased algorithmic predictions influenced human decision-makers through the anchoring effect, making them make more biased final decisions compared to the humans working without algorithmic predictions. In addition, the author showed that informing the human DMs with the AI's error rate reduced the errors in the humans' final decisions (Rhue, 2019).

Also focused on the anchoring effect of AI bias on human decision-makers' decision bias, Vaccaro (2019) showed that DMs working with a biased AI made more biased decisions compared to the DMs working alone. The author thus argued that in certain cases such as this where the anchoring effect is at play, the human-machine collaborative decision-making paradigm caused decision bias to worsen compared to the humans working by themselves. The two studies therefore raised the concerns that including AI in decision-making process alongside humans does not necessarily improve human decision-making performance in terms of bias, and can actually make the bias worse.

Although in Vaccaro's (2019) and Rhue's (2019) experiments, participants worked with an AI on the same tasks (predicting recidivism rate in the former, and guessing people's age and rating their beauty score in the latter) for a number of times, they did so in a one-shot manner.

More specifically, these participants working with an AI made their multiple decisions one immediately after another in a single decision period, and did not receive feedback about their own and the AI's decision performances, either after each decision or after a series of decisions throughout the experiment. Thus, these two studies did not consider the possible exposure effects of the DMs who worked with an AI and received performance feedback repeatedly for multiple decision periods over time.

In addition, for the comparison of performance regarding decision bias among the human-machine teams, the humans alone, and the machine alone, these two studies only employed one particular AI, thus largely overlooking AI's varying performance characteristics (varying levels of bias) that could potentially alter human DMs' behavior, perceptions, and decision-making performance when working with the AI.

Further, these studies did not examine the impact of working with a biased AI on the decision-makers' perception of AI fairness, which potentially plays an important role in explaining the humans' decision-making behavior, their performance and decision bias resulting from their collaboration with the AI.

3. Theoretical Framework

We use insights from computer science/human-computer interaction, statistics, economics, along with Rational Choice Theory as the theoretical basis for our research model. In the following subsections, we review the multiple approaches to defining and measuring decision-making bias, along with extant empirical findings on people's perceptions of fairness. We then present our research hypotheses grounding in Rational Choice Theory and previous empirical findings.

3.1 Bias and Fairness

We first clarify what we refer to when we use the terms bias and fairness in our paper. We view bias as an objective construct that can be mathematically measured, and fairness as a subjective, perception-based construct, used for evaluating whether a decision is in accordance with established ethical values and social norms.

Bias, which in our study we specifically focus on group bias, arises when there is a measurable difference in decision outcomes, through comparing the predictive accuracy and/or error rate balance between groups of people who are affected by the decisions. The source of this bias can come from the data used for training the predictive algorithm (i.e., data bias), the actual coding of algorithms (i.e., algorithm bias), or human decision-makers' minds (i.e., cognitive bias). As such, bias can be objectively (mathematically) established by inspecting the data, the predictions of the algorithms, or the decisions of the humans or human-machine teams.

Fairness, which in our study we also specifically focus on group fairness, on the other hand is when people perceptually evaluate whether the differences found in decision outcomes are fair or not for all of the groups that the decisions affect. In other words, fairness is a higher level evaluation of the implications of decision bias across groups. Fairness deals with social norms about justice and equal outcomes being viewed as the desired outcomes. More specifically, fairness is predicated on the philosophical assumptions that society needs to create equal outcomes across different groups of people.

To that end, we use bias in our study to specifically refer to the objective, mathematical measures of whether or not a set of predictions or decisions produce unequal outcomes among groups of people. We use fairness to specifically refer to people's subjective evaluations of whether or not a set of predictions/decisions complies with social norms of justice and equal outcomes among groups.

3.2 Measuring Bias

The decision context of our study — loan applications review — falls under the binary classification decision problem. Many other common, complex decisions also belong to this category of decision-making type. Examples include hiring decisions, college admission decisions, medical diagnoses, predicting recidivism when setting bail amounts, and so on.

In this classification decision problem, there are typically two possible decisions/predictions and two possible outcomes. In the case of loan applications review, the two possible decisions/predictions would be approving a loan application and rejecting it. Because the variable of interest to lending organizations is usually whether the loan will be a bad loan (defaulting, late payments, etc.), approving a loan—predicting low risk of becoming a bad loan—is considered a negative decision, and rejecting a loan—predicting high risk of becoming a bad loan—is considered a positive decision. Similarly, the two possible outcomes of loan applications—turning out to be a good loan and turning out to be a bad loan would constitute respectively a negative and a positive outcome. A confusion matrix (see table 1) can thus be constructed to evaluate the different possible types of decision/prediction errors resulting from this decision problem.

For this type of classification decision problem, there are various statistical measures (e.g., predictive parity, accuracy equity, error rate balance, etc.) that are commonly used to assess whether an algorithm or a decision-maker is biased against certain groups of people. Depending on the decision context and the decision-maker's priority, one or several of them can be selected and compared among different groups of people.

The first set of measures are referred to as predictive parity among groups. This consists of positive predictive value and negative predictive value, which either or both are computed and compared among groups.

$$Positive\ Predictive\ Value = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Negative\ Predictive\ Value = \frac{True\ Negatives}{True\ Negatives + False\ Negatives}$$

The second set of measures are referred to as error rate balance among groups. This consists of false positive balance and false negative balance. Specifically, false positive rate and/or false negative rate are computed and then compared among groups.

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives +\ True\ Negatives}$$

$$False\ Negative\ Rate = \frac{False\ Negatives}{False\ Negatives + True\ Positives}$$

Third, accuracy equity is another measure that can be used to determine whether there is bias against certain groups. Specifically, the prediction/decision accuracy rate is computed and compared among groups.

$$Accuracy\ Rate = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + True\ Positives +\ False\ Positives}$$

In our paper, we focus on specifically two measures: False Positive Balance (difference in the false positive rate between groups) and False Negative Balance (difference in the false negative error rate between groups).

3.3 Evaluating Fairness

Even though the statistical measures described above, among many others, are considered objective, mathematical measures of bias, there is usually a significant amount of subjectivity involved in choosing which measure(s) to assess decision fairness. In fact, different stakeholders have been found to often have differing opinions about which objective measure(s) should be used to evaluate whether a set of decisions or predictions are fair or not (Cowgill & Tucker, 2020).[2] For instance, in the COMPAS (an algorithm predicting recidivism risk) case, the defendants and social justice critics were more interested in false positive balance between racial groups, whereas the designers of the algorithm (Northpointe) were more interested in predictive parity.

Similarly, empirical research on perceptions of fairness have also demonstrated that people's fairness perceptions vary drastically among different stakeholders. For instance, in the human-computer interaction literature, Lee and Baykal (2017) showed that despite the programmers' efforts to design unbiased algorithms, a sizable portion of users felt they were treated unfairly by the AI. The authors found that the main reason for this was fairness definitions differed between the end users and the algorithm designers/programmers, and even among the end users.

Relatedly, in the economics literature, Babcock et al. (1995) demonstrated through an experiment that in bargaining decision contexts, the people involved judged fairness differently in a self-serving manner, and this tendency intensified as the amount of provided information increased. Their results echo findings in the experimental psychology literature that people often view decisions that satisfy their self-interests as fairer than the ones that do not.

Taking a different approach, Konow (2009) examined if people can achieve convergence in fairness perceptions when they are (unlike in the cases and studies described thus far in this section) not the directly implicated parties in the decision outcomes. The author found that

[2] See also 21 Fairness Definitions and Their Politics, Narayanan ACM FAccT '18

consensus on perceptions of what is fair in this case did occur when more detailed information about the decision context was provided to the people evaluating decision fairness. Further, the author also found that varying personal characteristics did not significantly influence people's fairness perspectives (Konow, 2009).

4. Research Hypotheses

4.1 AI Bias and the Human-Machine Teams' Performance

Based on anchoring theory (Tversky & Kahneman, 1974), which has been demonstrated in empirical studies regarding the anchoring effect of AI predictions in human-machine collaborative decision-making (Vaccaro 2019, Rhue 2019), we posit that those working with the biased AI would be influenced to some extent by its predictions, and thus would make more biased decisions and in turn make less profit for their organization compared to those working with those working with an unbiased AI. Hence, our H1a - H1c :

H1a: *AI prediction bias is positively associated with human-machine teams' decision bias in terms of false positive rate imbalance.*

H1b: *AI prediction bias is positively associated positively associated with human-machine teams' decision bias in terms of false negative rate imbalance.*

H1c: *AI prediction bias is negatively associated with the organizational profit produced by the human-machine teams.*

4.2 AI Bias and the DMs' Perceptions

In our experiment, the participants are financially rewarded and penalized based on decision-making performance, with their incorrect decisions being quite seriously penalized. Thus, they can be considered a group of stakeholders directly implicated by the outcomes of their joint loan review decision-making process with the AI. This means that participants' perceptions of AI fairness, motivated by their own financial interests, would depend to some extent on the performance of the AI assigned to them, which as per H1 would likely influence the performance of their collaborative work with the AI. Because participants receive immediate feedback after each round in the experiment on the performance of the AI predictions and of their joint decision-making with the AI, along with how much they earned personally, the participants working with the biased AI should be more likely than those working with the unbiased AI to perceive their algorithmic partner as less fair. We base our postulation on the empirical findings regarding the anchoring effect of AI bias and perceived fairness reviewed above and on Rational Choice Theory (RCT) which would predict that DMs, independent of their personal biases if any, would want to maximize their earnings. Hence our H2 :

H2: *AI prediction bias is negatively associated with human decision-makers' perceived algorithmic fairness in human-machine collaboration.*

4.3 Exposure and Performance

Also following the logics of RCT and additionally the Horndal plant labor (Lundberg, 1961) effect found in the firm progress function literature, we posit that through repeatedly interacting with the AI and continually receiving feedback on the performance of the AI and of their joint work with the AI, decision-makers – both those working with the biased and unbiased AI, would begin to adapt their behavior, motivated by their self-interests to maximize their earnings, which in turn would improve their performance relative to the AI operating alone over time. Specifically, the difference in the profit they make for their organization relative to that of the AI would increase over time. Hence, our H3 :

H3: *There are positive exposure effects in human-machine collaboration: over time, the difference between the organizational profit made by the human-machine teams' and that made by the AI alone increases.* (Note : We focused on the difference in organizational profit to remove the potential influence (anchoring) of AI predictions on the human-machine teams' performance.)

5. Study Design

5.1. Research Method, Dataset, Algorithm Predictions

We adopt experimental economics as our main research method and design a decision-making platform that simulates actual organizational decision contexts. We have run experimental pilot sessions with 28 participants in the No Bias treatment and 35 in the Prediction Bias treatment. In addition, we also ran a session of the Prediction Bias treatment with extended decision time (Prediction Bias Extended treatment, n = 9) to preliminarily test whether allowing the DMs additional time for reviewing the loan applications would have an impact on their performance and decision fairness.

The participants were undergraduate students majoring in Business at a large university in the U.S. We paid participants with financial rewards to induce rational behavior (Smith, 1976). Participants earned a flat reward ($5) for completing the entire session and a bonus reward determined by their decision-making performance in the experiment (total ranging $5 – $20, averaging $10).

In the experiment, participants worked with an AI tool to evaluate 100 consumer loan applications and made the final decisions (approve/reject). We derived the 100 loan applications that the participants reviewed from a real world, large-scale, historic dataset from LendingClub (n = ~50,000). This 100-loan subset contains 50 loans that performed poorly (bad loans) and 50 loans that performed well (good loans).

This historic dataset includes real consumer loan applications' information (e.g., occupation, income, outstanding debt, loan purpose, loan requested amount, income, etc.) and real loan performance data for the loans that were funded (e.g., loan interest rates, monthly payments, defaults, late payments, etc.). Because of that, we knew which loans performed well and which did not. The participants in the experiment saw the simplified consumer loan application data but not the loan performance data when they made their decisions. The participants' decision-making performance, however, was determined based on the real loan performance data.

5.2 Experimental Task

The experimental task comprised 10 repeated but distinct rounds. In each, the DMs made decisions (approve/reject) for 10 loan applications (total of 100) with the aid of an AI. The DMs received immediate decision feedback at the end of each round. There were 3 stages in each round: *Initial Decisions, Final Decisions,* and *Results*. In the Initial Decisions stage, the DMs viewed a loan data table that contained the loan applicants' information, made their initial decisions (approve/reject), and rated their initial level of confidence (1-100). In the Final Decisions stage, the DMs viewed the same loan table but that had been updated with their own initial decisions and the AI predictions (presented next to each other). Based on this new information, the DMs now had the opportunity of revising their initial decisions and confidence levels before submitting the final decisions and confidence ratings. In the Results stage, the DMs received feedback on how the loans actually performed, the DMs' and the AI's performance statistics (i.e., rate of correct decisions), and their cash earnings from each decision, the round total, as well as the cumulative total.

5.3 Financial Reward Structure

The DMs in all experimental conditions earned $0.2 for each correct decision (i.e., approve or reject), and $0 for each wrong reject (false positve). They lost $0.4 for each wrong approve decision (false negative). This incentive mechanism was designed to connect the DMs' rewards to the assumed risk-averse, conservative lending strategy that we stipulated for the study, i.e., discouraging the DMs from making risky bets via high penalties for the wrong approve decisions (false negative). Table 2 shows this reward structure in the form of a pay-off matrix.

We induced this assumed, fixed conservative, risk-averse organizational lending strategy in our experiment, by telling the subjects in the instructions to "imagine that you are a loan officer for a traditional retail bank" and "your task is to identify good loan applications for approval and risky loan applications for rejection".

5.4 Experimental Procedures

The experimental sessions were conducted synchronously online via Zoom meetings, each lasting between 60–75 minutes from check-in to check-out. The number of participants varied (ranging 2–16) for each session. In each session, after all participants have joined the Zoom meeting, the experimenter welcomed and introduced briefly to the participants the basic structure of the session which included 3 parts: a pre-experiment survey, a decision-making task, and a post-experiment survey.

Next, the experimenter sent the participants the URL to the online experimental platform and instructed the participants to complete the pre-experiment survey that consists of questions about their demographic characteristics and personality traits. After making sure every participant had finished the survey, the experimenter presented the instructions slides on the Zoom meeting screen and also at the same time read the instructions aloud to explain in detail the loan review task and the financial reward structure, i.e., the pay-off matrix detailing how much they will earn or lose for each correct (incorrect) loan decision.

The participants did not know that there were multiple treatments, nor did they know which particular treatment they were participating in. After the participants finished the experimental task, they were required to complete the post-experiment survey which contained questions about their decision-making process and their perceptions (e.g. regarding how fair the AI was, how much they trusted the AI, how fair their cash earnings were, etc.). Finally, the participants were dismissed and later on were paid via online bank transfers according to how much money they earned in the task.

5.5 Treatment Manipulations

5.5.1 The Two Groups of Loan Applicants

We operationalized the two groups of loan applicants in our experiment by randomly splitting the 100 historic loan applications used in our experiment into two (fictional) groups with equal size: Purple (n = 50) and Orange (n = 50).

We represented these two groups in the experiment by highlighting the loans' IDs and group names respectively in the (Loan) ID and Applicant Type columns in the loan applications table shown to the DMs, accordingly with two colors – purple and orange (See Figure 1 for a partial view of the experimental interface). We also informed the participants of this detail in the instructions portion at the start of the experiment.

We operationalized the two loan applicant groups through color labels (Orange and Purple) because we wanted to avoid any confounding effects that might result from using the applicants' personal information such as gender, ethnicity, occupation, names, or any other stimuli (visuals, names, etc.) which different people can have different personal responses to. We are not interested in how people personally respond to certain groups of the population.

In both treatments, there are 25 loans for each of the four categories: Good Purple, Bad Purple, Good Orange, Bad Orange.

5.5.2 Prediction Bias

In the Prediction Bias treatment, while the loan data used in the experiment are unbiased with respects to the Orange and Purple groups, the algorithm predictions shown to the DMs are biased against the Purple group and for the Orange group. Specifically, the Orange and Purple groups have different false positive rates (Orange = 0.2, Purple = 0.6) and different false negative rates (Orange = 0.6, Purple = 0.2). In other words, the algorithm predictions in the Predict Bias treatment inaccurately discriminate against the Purple (rejecting more good loans) compared to the Orange group, while inaccurately favoring the Orange (approving more bad loans) over the Purple group. The overall accuracy of the algorithm predictions in the Prediction Bias condition is 60%.

In the No Bias condition, both the loan data and the algorithm predictions used in the experiment are unbiased between the two groups. Specifically, both Orange and Purple groups have the same rates of 0.2 for false positive and false negative. The overall accuracy of the algorithm predictions in the No Bias condition is 80%.

Tables 3 – 5 summarize the AI's decision bias metrics in each condition.

5.5.3 Decision Time

In the Prediction Bias Extended treatment, the Initial Decision stage had a time limit of 3 minutes, the Final Decision stage 2 minutes, and the Results stage 1.5 minutes, whereas in the other two treatments run with regular decision time, the Initial Decision stage had a time limit of 2 minutes, the Final Decision stage 1 minute, and the Results stage 1 minute as well.

6. Results

6.1 Prediction Bias and Human-Machine Teams' Performance (H1a-c)

To examine the impact of the experiment factor (Prediction Bias) on the human-machine teams' performance, in terms of decision bias (false positive rate imbalance between Purple and Orange groups, false positive rate imbalance between Orange and Purple groups), and organizational profit, we developed the following regressions:

$$FPI^{Pu-Or} = \beta_0 + \beta_{PER_i}\Sigma_{i=1}^{9}PER_i + \beta_{PB}PB + controls + \varphi$$

$$FNI^{Or-Pu} = \gamma_0 + \gamma_{PER_i}\Sigma_{i=1}^{9}PER_i + \gamma_{PB}PB + controls + \mu$$

$$OP = \alpha_0 + \alpha_{PER_i}\Sigma_{i=1}^{9}PER_i + \alpha_{PB}PB + controls + \varepsilon$$

In the equations above, FPI^{Pu-Or}, FNI^{Or-Pu}, OP respectively refer to false positive imbalance between Purple and Orange groups, false negative imbalance between Orange and Purple groups, and organizational profit; α, β, γ are the intercepts; PER_i (i = 1 to 9) are dummy variables that represent the 10 decision rounds; PB refers to prediction bias (1 for bias, 0 for no bias); and $\varepsilon, \varphi, \mu$ are error terms.

The dummy variables representing the 10 decision rounds are included in the regressions because we collected decision-making data over 10 repeated and distinct rounds, with feedback provided to the participants after each round, and thus we wanted to control for the possible learning effects that we conjectured had occurred over the rounds.

The final, unbalanced sample used for analyzing our two main experimental treatments (No Bias and Prediction Bias) included 63 participants. Because we used round-level data, we had in total 10 rounds × 63 participants = 630 observations. Table 6 presents the descriptive analysis of our experimental data.

The other control variables account for the participants' traits and demographics which include quantitative and financial competency, college level, ethnicity, gender, and cognitive style. The descriptive statistics of the participants' performance, traits, demographics, and perceptions are included in Tables 7, 8.

Specifically, we conducted a hierarchical multiple linear regression for each of the performance variables, in 3 stages (see Table 9a-c, Models 0–2). In each, we respectively added the main independent variable (prediction bias), the experimental period dummy variables PER_i (i =1–9), and finally in Model 2, the other control variables.

We computed false positive imbalance between purple and orange by computing the difference in the false positive error rates between the purple group and the orange group (subtracting the false positive error rate of orange group from that of purple group).

We computed false negative imbalance between orange and purple by computing the difference in the false negative error rates between the orange group and the purple group (subtracting the false negative error rate of purple group from that of orange group).

We computed organizational profit (in USD) resulted from the loan review decisions by calculating the net present values of the historical, approved loans.

Results show that the human-machine teams in the biased predictions treatment had a significantly higher positive rate imbalance between purple and orange groups than those in the unbiased treatment (8% higher, $p < 0.05$). In addition, working with biased predictions led to significantly higher negative rate imbalance between orange and purple groups compared to working with unbiased predictions (10.4%, $p < 0.001$). In other words, the presence of biased algorithmic predictions did negatively influence (anchor) the human DMs' decisions in the same way that the AI is biased: discriminating against the purple group and favoring the orange group.

The human-machine teams in the biased predictions treatment also made significantly less profit for the organization than those in the unbiased predictions treatment, costing their organization around $6,036 on average in each round ($p < 0.001$). Thus, our H1a, H1b, and H1c are supported.

6.2 Prediction Bias and Human DMs' Algorithmic Fairness Perception (H2)

We measured perceived algorithmic fairness by asking the participants after they had completed the entire experimental task to rate on a Likert scale of 1-7 (1 = Strongly Disagree, 7 = Strongly Agree) regarding whether they think that the AI treated all the loan applicants equally.

To examine the impact of prediction bias on perceived algorithmic fairness, we conducted a two-way t-test to compare the perceived algorithmic fairness measure between the biased predictions treatment and the unbiased treatment. Results show that there was a small and insignificant ($p = 0.4$) difference in perceived fairness between those working with the biased AI ($m = 4.8$) those working with the unbiased AI ($m = 5.1$). We further found that, however, the former rated their trust in the AI ($m = 11.3$) significantly ($p < 0.01$) less than the latter ($m = 13.3$) for a decrease of 15 %. We measured this AI Trust construct by adapting the 3-item Trust Belief scale by Robert, Dennis, and Hung (2009). Thus, the DMs appeared to implicitly recognize the biased predictions of the AI rather than explicitly.

Interestingly, we found a significant, albeit weakly, difference ($p = 0.08$) when we compared perceived algorithmic fairness between the DMs working with the biased predictions and given extra time limits for making decisions ($m = 4.1$) and those working with the unbiased predictions ($m = 5.1$), for a decrease of 19.6 %.

6.3 Exposure and Organizational Profit (H3)

We first computed the difference in organizational profit between the human-machine teams and the AI alone. Then, following the standard practice in the experimental economics literature (Cadsby & Maynes, 1998; Embrey et al., 2018; Fréchette, 2009) for analyzing potential learning effects in repeated round experiments, we compared the average of this measure between clusters of decision periods. Specifically, we looked at the earlier 5 rounds vs. the later 5 rounds. Overall, across all treatments of the human-machine teams, the average difference in organizational profit compared to the AI alone's organizational profit per round increased significantly in the later periods, with a rise of of 179% ($p < 0.001$). More specifically, in the earlier 5 rounds, the human-machine teams made on average $3,012 less than the AI alone while in the later 5 rounds, they made on average $2,383 more than the AI alone.

Analyzing conditions separately, we found that this improvement in organizational profit relative to the AI alone was also significant. However, those working with the biased AI improved in the later 5 rounds considerably more than those working with the unbiased AI. Specifically, while the former improved in their organizational profit relative to the AI alone by 383%, the latter only did so by 59%. Moreover, while those working with the biased AI in the later 5 rounds made for the organization on average $4,546 more than the AI alone in each round, those working with the unbiased AI in the later 5 rounds made for the organization on average $2,395 less than the AI alone in each round.

Intrigued by this finding, we further directly compared the human-machine teams' performance with the AI alone's performance using the Wilcoxon Signed Rank test for matched pairs. We found that overall, across all treatments, on average in each round the human-machine teams outperformed the AI alone, in both achieving higher organizational profit ($p = 0.06$) and reducing bias – lowering both false positive imbalance ($p < 0.001$) and false negative imbalance ($p < 0.001$). This superiority was found to be stronger when we only included the performance measures of the later 5 rounds. Analyzing treatments separately, we found that, however, human-machine teams in the unbiased treatment in particular did not outperform the unbiased AI alone.

7. Discussion

We contribute to research on bias in human-machine collaboration by showing that, contrary to what current related studies have shown, human DMs can learn to work with an imperfect, biased AI to improve significantly their performance over time and also outperform the biased AI working on its own. Working with a biased AI is often the case in reality as it is almost impossible to design a completely unbiased AI. As such, our research shows that with repeated interactions, timely feedback, and an appropriate incentive mechanism, organizations can reap benefit from having human DMs work with biased algorithms, to reduce bias significantly and improve their profit.

Second, we contribute by empirically demonstrating that human DMs working with a less biased (in this case, a completely unbiased) AI can produce less biased decisions and made higher organizational profit. This shows that bias mitigation in the algorithm component can translate to bias mitigation in human-machine decision-making partnership when combined with the right conditions – incentive mechanism, time, exposure, and clear feedback. Most current studies have not examined or demonstrated this.

Our third contribution is adding a dynamic perspective. Unlike most present behavioral studies on bias in human-machine collaboration, our experiment features the repeated round design (as opposed to one-shot), which allowed us to perform a longitudinal analysis that shows the more nuanced dynamics of how DMs interact with AI over time. In real-world organizational decision-making contexts, it is usually through repeated interactions (exposure) with new technologies that organizaitonal DMs gain familiarity and improve their performance significantly.

Methodologically, we are the first, to our knowledge, to examine bias in human-machine decision-making that operationalized bias in terms of error rate imbalance between two fictional groups with neutral labels, which allowed us to reduce (if not completely avoid) the potential confounding effects of DMs' personal biases.

On a related note, we expect to find similar results in future research where instead of manipulating the loan applicant groups through fictional neutral labels, we do so by using certain information such as race, gender, religion, etc., which DMs likely have varying personal responses (biases) to. Specifically, we expect that over time, DMs motivated by their self-interest to maximize their financial earnings, can also learn to adapt their decision-making behvaior, and to some extent overcome both their own biases and the AI's biases to improve their performance, reduce decision bias, and increase profit for their organization.

References

Adomavicius, G., & Yang, M. (2019). Integrating Behavioral, Economic, and Technical Insights to Address Algorithmic Bias: Challenges and Opportunities for IS Research. *SSRN*.

Autor, D., Mindell, D., & Reynolds, E. B. (2019). *The Work of the Future: Shaping Technology and Institutions*. Massachusetts Institute of Technology.

Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review*, *85*(5), 1337–1343.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013.

Cowgill, B., & Tucker, C. E. (2020). Economics, fairness and algorithmic bias. *Working paper.*

Daugherty, P. R., & Wilson, H. J. (2018). *Human+ Machine: Reimagining Work in the Age of Ai*. Harvard Business Press.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dutton, J.M. and Thomas, A. Treating progress functions as a managerial opportunity. *Academy of Management Review*, *9*, 2 (1984), 235–247.

Kasy, M., & Abebe, R. (2020). *Fairness, equality, and power in algorithmic decision making*. Working paper.

Khademi, A., Lee, S., Foley, D., & Honavar, V. (2019). Fairness in algorithmic decision making: An excursion through the lens of causality. *The World Wide Web Conference*, 2907–2914.

Konow, J. (2009). Is fairness in the eye of the beholder? An impartial spectator analysis of justice. *Social Choice and Welfare*, *33*(1), 101-127.

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. Discussion-based social division. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1035–1048.

Lu, T., Zhang, Y., & Li, B. (2019). *The Value of Alternative Data in Credit Risk Prediction: Evidence from a Large Field Experiment*.

Lundberg, E. *Produktivitet Och Räntabilitet: Studier i Kapitalets Betydelse Inom Svenskt Näringsliv*. Studieförbundet Näringsliv och samhälle, 1961.

Rhue, L. (2019). Beauty is in the AI of the Beholder: How Artificial Intelligence Anchors Human Decisions on Subjective vs. Objective Measures. *Proceedings of the International Conference on Information Systems*.

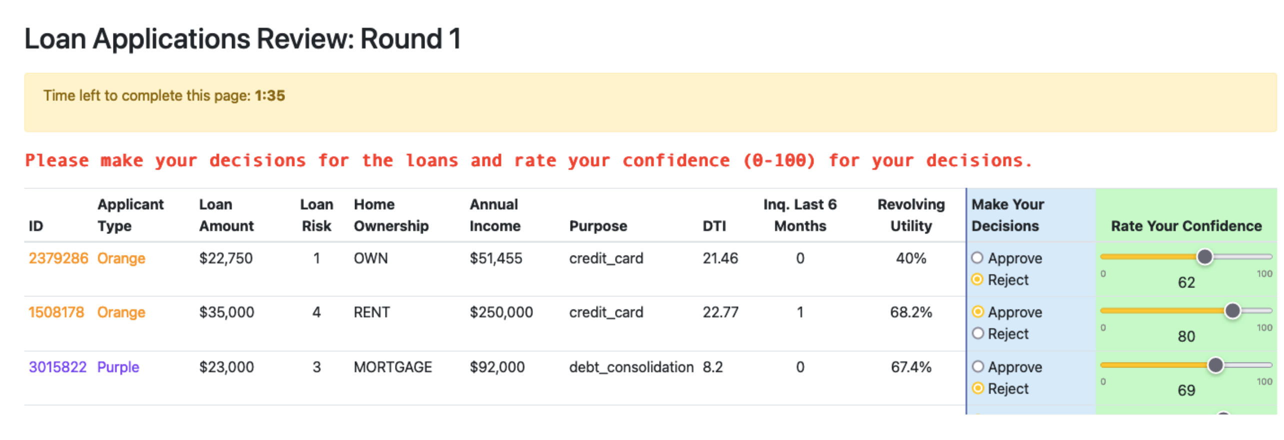
Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, *66*(2), 274–279.

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

Vaccaro, Michelle Anna. (2019). Algorithms in Human Decision-Making: A Case Study With the COMPAS Risk Assessment Software. Bachelor's thesis, Harvard College.

Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., & Cui, W. (2020). Algorithmic Decision Making with Conditional Fairness. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2125–2135.

Figures and Tables

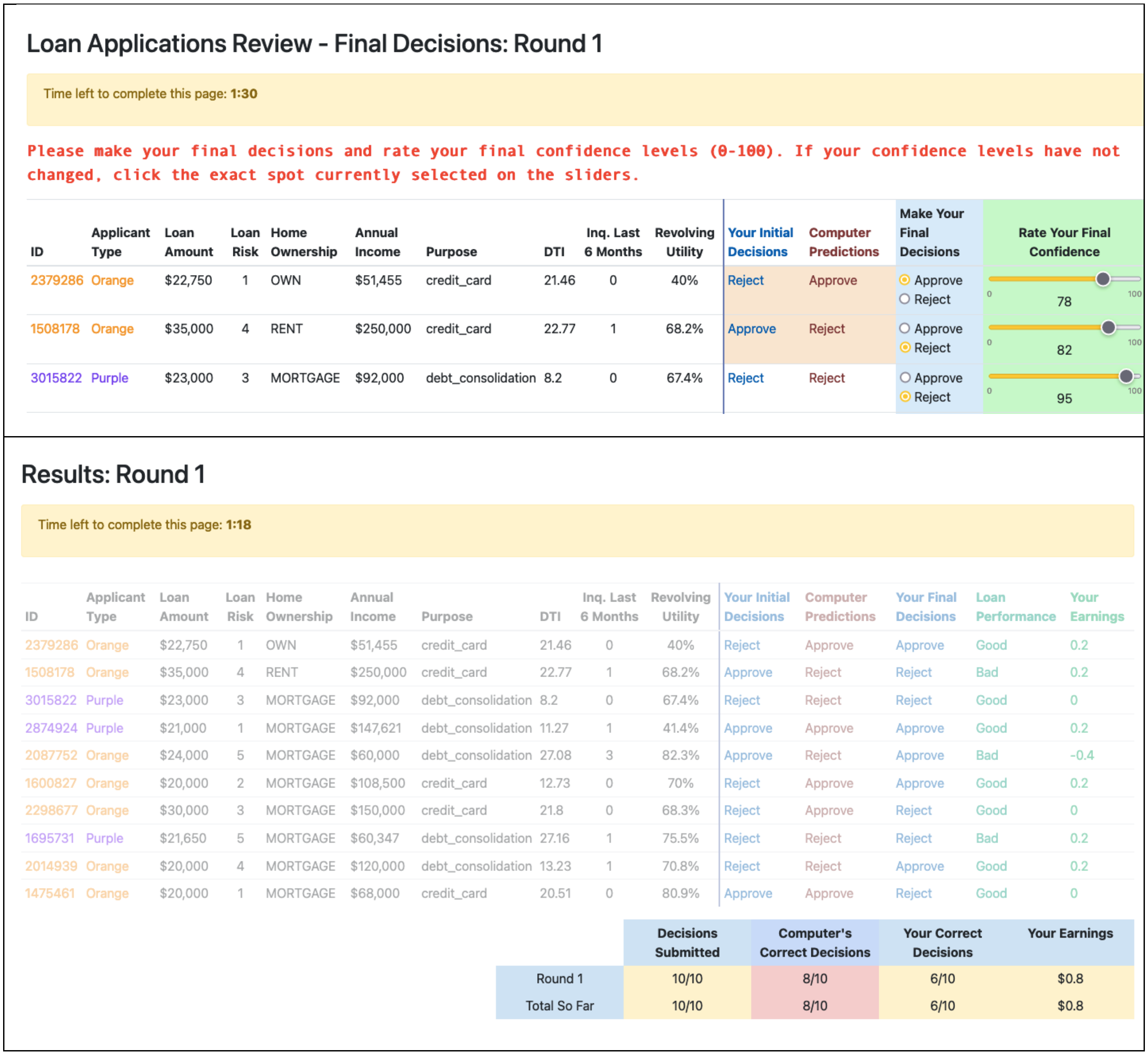


Loan Applications Review: Round 1

Time left to complete this page: **1:35**

Please make your decisions for the loans and rate your confidence (0-100) for your decisions.

ID	Applicant Type	Loan Amount	Loan Risk	Home Ownership	Annual Income	Purpose	DTI	Inq. Last 6 Months	Revolving Utility	Make Your Decisions	Rate Your Confidence
2379286	Orange	$22,750	1	OWN	$51,455	credit_card	21.46	0	40%	○ Approve ● Reject	0 100 62
1508178	Orange	$35,000	4	RENT	$250,000	credit_card	22.77	1	68.2%	● Approve ○ Reject	0 100 80
3015822	Purple	$23,000	3	MORTGAGE	$92,000	debt_consolidation	8.2	0	67.4%	○ Approve ● Reject	0 100 69



Loan Applications Review - Final Decisions: Round 1

Time left to complete this page: **1:30**

Please make your final decisions and rate your final confidence levels (0-100). If your confidence levels have not changed, click the exact spot currently selected on the sliders.

ID	Applicant Type	Loan Amount	Loan Risk	Home Ownership	Annual Income	Purpose	DTI	Inq. Last 6 Months	Revolving Utility	Your Initial Decisions	Computer Predictions	Make Your Final Decisions	Rate Your Final Confidence
2379286	Orange	$22,750	1	OWN	$51,455	credit_card	21.46	0	40%	Reject	Approve	● Approve ○ Reject	0 100 78
1508178	Orange	$35,000	4	RENT	$250,000	credit_card	22.77	1	68.2%	Approve	Reject	○ Approve ● Reject	0 100 82
3015822	Purple	$23,000	3	MORTGAGE	$92,000	debt_consolidation	8.2	0	67.4%	Reject	Reject	○ Approve ● Reject	0 100 95

Results: Round 1

Time left to complete this page: **1:18**

ID	Applicant Type	Loan Amount	Loan Risk	Home Ownership	Annual Income	Purpose	DTI	Inq. Last 6 Months	Revolving Utility	Your Initial Decisions	Computer Predictions	Your Final Decisions	Loan Performance	Your Earnings
2379286	Orange	$22,750	1	OWN	$51,455	credit_card	21.46	0	40%	Reject	Approve	Approve	Good	0.2
1508178	Orange	$35,000	4	RENT	$250,000	credit_card	22.77	1	68.2%	Approve	Reject	Reject	Bad	0.2
3015822	Purple	$23,000	3	MORTGAGE	$92,000	debt_consolidation	8.2	0	67.4%	Reject	Reject	Reject	Good	0
2874924	Purple	$21,000	1	MORTGAGE	$147,621	debt_consolidation	11.27	1	41.4%	Approve	Approve	Approve	Good	0.2
2087752	Orange	$24,000	5	MORTGAGE	$60,000	debt_consolidation	27.08	3	82.3%	Approve	Reject	Approve	Bad	-0.4
1600827	Orange	$20,000	2	MORTGAGE	$108,500	credit_card	12.73	0	70%	Reject	Approve	Approve	Good	0.2
2298677	Orange	$30,000	3	MORTGAGE	$150,000	credit_card	21.8	0	68.3%	Reject	Approve	Reject	Good	0
1695731	Purple	$21,650	5	MORTGAGE	$60,347	debt_consolidation	27.16	1	75.5%	Reject	Reject	Reject	Bad	0.2
2014939	Orange	$20,000	4	MORTGAGE	$120,000	debt_consolidation	13.23	1	70.8%	Reject	Reject	Approve	Good	0.2
1475461	Orange	$20,000	1	MORTGAGE	$68,000	credit_card	20.51	0	80.9%	Approve	Approve	Reject	Good	0

	Decisions Submitted	Computer's Correct Decisions	Your Correct Decisions	Your Earnings
Round 1	10/10	8/10	6/10	$0.8
Total So Far	10/10	8/10	6/10	$0.8

Figure 1. Top to Bottom : Initial Decision Page (partial view), Final Decision Page (partial view), Results Page

Table 1. Classification Decision Problem – Confusion Matrix

	Approve (Predicting Low Risk of Bad Outcome)	Reject (Predicting High Risk of Bad Outcome)
Good Loan Outcome	True Negative	False Positive
Bad Loan Outcome	False Negative	True Positive

Table 2. Participants' Pay-off Matrix in the Experiment

	Approve	Reject
Good Loan Outcome	$ 0.2	$ 0
Bad Loan Outcome	- $0.4	$ 0.2

Table 3. AI's Confusion Matrix - Prediction Bias Treatment (Against Purple, For Orange)

	AI Approves	AI Rejects
Good Loans	30 True Negatives 10 Purple, 20 Orange	20 Type I Errors 15 Purple, 5 Orange
Bad Loans	20 Type II Errors 5 Purple, 15 Orange	30 True Positives 20 Purple, 10 Orange

Table 4. AI's Confusion Matrix – No Bias Treatment (Orange and Purple Treated Equally)

	AI Approves	AI Rejects
Good Loans	40 (True Negatives) 20 Purple, 20 Orange	10 (False Positives) 5 Purple, 5 Orange
Bad Loans	10 (False Negatives) 5 Purple, 5 Orange	40 (True Positives) 20 Purple, 20 Orange

Table 5. AI's False Positive and False Negative Rates Across Treatments

	Treatments	
	No Bias	Prediction Bias
False Positive Rate	All Loans: 0.2; Orange: 0.2; Purple: 0.2	All Loans: 0.4; Orange: 0.2; Purple: 0.6
False Negative Rate	All Loans: 0.2; Orange: 0.2; Purple: 0.2	All Loans: 0.4; Orange: 0.6; Purple: 0.2

Table 6. Descriptive Statistics of Performance Measures

	Prediction Bias	No Bias
Organizational Profit	-$22,929.47 ($18,828.66)	-$17,164.55 ($16,627.34)
False Positive Imbalance (Purple – Orange)	0.14 (0.43)	0.06 (0.46)
False Negative Imbalance (Orange – Purple)	0.16 (0.43)	0.05 (0.37)
Purple Type I Rate	0.34 (0.35)	0.30 (0.36)
Orange Type II Rate	0.40 (0.35)	0.26 (0.28)
Type I Rate	0.28 (0.22)	0.27 (0.23)
Type II Rate	0.33 (0.22)	0.24 (0.18)

Note: Numbers outside parentheses are the means, inside are the standard deviations

Table 7. Participants' Ethnicities

East Asian	17	23.6%
South Asian	5	6.9%

South East Asian	7	9.7%
Black	5	6.9%
Hispanic/Latino	14	19.4%
Pacific Islander	1	1.4%
White - Europe	13	18.1%
White - Middle East or North Africa	10	13.9%

Table 8. Participants' Demographics

Gender	Female : 22 (30.6%); Male: 50(69.4%)
College Level	Sophomore: 41 (56.9%); Junior: 22 (30.6%); Senior : 9 (12.5%)
	Upper (Senior/Junior): 31 (43.1%) Lower (Freshman/Sophomore) : 41 (56.9%)
Minority - Ethnicity [3]	30 (41.7%)
Quant/Financial Competence	Mean (SD) : 48.6 (19) ; Min : 0 ; Med : 51.7 ; Max : 85.1
Cognitive Style – Intuition (vs. Analytic)	Mean (SD) : 25 (5.7) ; Min : 9 ; Med : 25 ; Max: 35

Table 9a. Impact of Prediction Bias on Organizational Profit

	Org Profit		
	Model 0	Model 1	Model 2
Prediction Bias	-5,764.927****	-5,764.927****	-6,035.511****
round_1		19,925.210****	19,925.210****
round_2		9,763.642****	9,763.642****
round_3		3,944.116*	3,944.116*
round_4		-15,592.470****	-15,592.470****
round_5		-3,527.116	-3,527.116
round_6		-3,405.232	-3,405.232
round_7		23,390.640****	23,390.640****
round_8		-15,438.520****	-15,438.520****
round_9		-6,872.970***	-6,872.970***
QuantFinComposite			92.732****
CollegeUpper			3,866.205****
EthnMinor			861.158
Male			277.726
FaithIntuition			-94.639
Constant	-17,164.550****	-18,383.280****	-22,523.610****
Observations	630	630	630
R^2	0.025	0.512	0.532
Adjusted R^2	0.024	0.504	0.521
F Statistic	16.164**** (df = 1; 628)	64.989**** (df = 10; 619)	46.550**** (df = 15; 614)
Note:			* p<0.1; ** p<0.05; *** p<0.01; **** p<0.001

[3] Includes Pacific Islander, Hispanic/Latino, Black, White – Middle East/North Africa

Table 9b. Impact of Prediction Bias on False Positive Imbalance

	False Positive Imbalance (Purple – Orange)		
	Model 0	Model 1	Model 2
Prediction Bias	0.081**	0.081**	0.080**
round_1		-0.484****	-0.484****
round_2		-0.749****	-0.749****
round_3		-0.754****	-0.754****
round_4		-0.798****	-0.798****
round_5			
round_6		-0.435****	-0.435****
round_7		-0.772****	-0.772****
round_8			
round_9		-0.725****	-0.725****
QuantFinComposite			-0.0003
CollegeUpper			-0.024
EthnMinor			-0.011
Male			-0.005
FaithIntuition			-0.006*
Constant	0.064**	0.654****	0.844****
Observations	504	504	504
R^2	0.008	0.342	0.347
Adjusted R^2	0.006	0.331	0.329
F Statistic	4.089** (df = 1; 502)	32.158**** (df = 8; 495)	19.992**** (df = 13; 490)
Note:		* p<0.1; ** p<0.05; *** p<0.01; **** p<0.001	

Table 9c. Impact of Prediction Bias on False Negative Imbalance

	False Negative Imbalance (Orange – Purple)		
	Model 0	Model 1	Model 2
Prediction Bias	0.104***	0.104****	0.104****
round_1		-0.209****	-0.209****
round_2		-0.630****	-0.630****
round_3		-0.370****	-0.370****
round_4		-0.058	-0.058
round_5		-0.511****	-0.511****
round_6		-0.526****	-0.526****
round_7		0.069	0.069
round_8		-0.045	-0.045
round_9		-0.209****	-0.209****
QuantFinComposite			-0.001
CollegeUpper			0.021
EthnMinor			-0.065**
Male			0.025

FaithIntuition			-0.001
Constant	0.053**	0.302****	0.381****
Observations	630	630	630
R^2	0.016	0.345	0.355
Adjusted R^2	0.014	0.334	0.339
F Statistic	10.136*** (df = 1; 628)	32.586**** (df = 10; 619)	22.538**** (df = 15; 614)

Note: * p<0.1; ** p<0.05; *** p<0.01; **** p<0.001