Kevin Bauer | Andrej Gill

Mirror, Mirror on the Wall: Machine Predictions and Self-Fulfilling Prophecies

SAFE Working Paper No. 313

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

Mirror, Mirror on the Wall:
Machine Predictions and Self-fulfilling Prophecies

Kevin Bauer,[1] Andrej Gill[2]

April 14, 2021

Abstract

We show that disclosing machine predictions to affected parties can trigger self-fulfilling prophecies. In an investment game, we experimentally vary investors' and recipients' access to a machine prediction about recipients' likelihood to pay back an investment. Recipients who privately learn about an incorrect machine prediction alter their behavior in the direction of the prediction. Furthermore, when recipients learn that an investor has disregarded a machine prediction of no-repayment, this further lowers the repayment amount. We interpret these findings as evidence that transparency regarding machine predictions can alter recipients' beliefs about what kind of person they are and what investors expect of them. Our results indicate that providing increased access to machine predictions as an isolated measure to alleviate accountability concerns may have unintended negative consequences for organizations by possibly changing customer behavior.

JEL classification: C91, D80, D91, O33

Keywords: Algorithmic transparency, algorithmic decision support, human–machine interaction

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

[2] Gutenberg School of Management and Economics, Johannes Gutenberg University Mainz, Jakob-Welder-Weg 4, 55128 Mainz, Germany, E-mail: gill@uni-mainz.de.

° We gratefully acknowledge research support from the Leibniz Institute for Financial Research SAFE and the Gutenberg School of Management and Economics, Johannes Gutenberg University Mainz.

1 Introduction

Machine generated predictions augment human decision making processes in a wide variety of domains. Examples include machine-based recidivism predictions used by judges to set bail (Kleinberg et al., 2018), candidate performance predictions used by HR managers to make hiring decisions (Horton, 2017; Hoffman et al., 2018), and credit scores used by loan officers to decide on credit applications (Huang et al., 2007; Kshetri, 2016). The rationale for relying on machine predictions in organizations is that they are considered more accurate and scalable than human predictions and thus more economically efficient (Brynjolfsson & Mcafee, 2017; Rahwan et al., 2019).

One tacit assumption behind our increasing reliance on machine predictions is that this reliance does not affect the behavior machines try to forecast. For instance, loan officers implicitly assume that using credit scores to assess creditworthiness does not influence applicants' actual repayment behavior. Similarly, managers expect that machine predictions about future performance will have no impact on applicants' real productivity. Naturally, this is a sound assumption when the individuals subject to a prediction (targets) are not aware of its use by decision-makers (users). However, the effects that may occur when targets become aware of machine predictions and their use largely remains an open question.

This paper aims to shed light on this issue. We examine how target behavior is impacted by the disclosure that their behavior is subject to machine prediction, and explore associated heterogeneity in predictive accuracy. As an illustration of the scenarios we have in mind, consider an individual who wants to apply for a loan. In preparation for her application, she requests (or purchases) access to her machine-produced credit scores from a service provider such as FICO or SCHUFA.[1] After learning about her predicted creditworthiness, she applies for a loan at a local bank. Due to legal requirements, e.g., Europe's General Data Protection

[1] For example see the websites `https://www.myfico.com` and `https://www.meineschufa.de/index.php` providing access to one's personal credit scores.

Regulation,[2] the bank has to disclose that the credit approval process involves checking her credit scores. Hence, the machine prediction is available to both the loan officer and the applicant, with the latter also being aware of the loan officer's access to the prediction. In this scenario we are interested in the following questions: Does learning about her predicted creditworthiness (and its use by the bank) affect the applicant's subsequent repayment behavior? What role does the accuracy of the prediction play? In summary, we are interested in the consequences of information transparency when it comes to machine predictions.

Several challenges arise when trying to examine the potential consequences of disclosing machine predictions. First, any impact on preferences is likely constrained by reputational concerns stemming from the repeated game nature of real life scenarios. Second, machine learning systems designed to make forecasts about people are necessarily unique, non-random assessments (e.g., due to organizational efficiencies, and legal requirements). Third, decision-makers' choice to override or follow a machine prediction is highly endogenous, depending on a variety of factors (e.g., reputational concerns, reliance on predictions). To address these challenges, we design a novel revealed-preferences experiment that we implement as an online study.

In our experiment, participants engage in three subsequent one-shot investment games (Berg et al., 1995). Investors initially choose between keeping or investing 10 monetary units (MU) with recipients, who, in the case of investment, decide how much of the tripled amount to repay to investors. The three investment games differ solely with a view to whether (i) no one, (ii) only the investor, or (iii) both the investor and recipient have access to a machine prediction about whether the recipient is most likely to pay back more than 10 MU (repayment) or not (no repayment). We employ the strategy method, allowing us to observe recipients' behavior for both actual and counterfactual predictions.

We report two main findings. First, privately informing recipients of the machine predictions made about them triggers self-fulfilling prophecies. On an aggregate level, when

[2] See Parliament & Council of European Union (2016)

recipients privately learn that the machine predicts they will not repay an investment, their repayment amount decreases by 20 percent relative to the baseline of no prediction. This decline is driven by participants who repay relatively large amount in the baseline. While we do not find aggregate level effects for privately disclosing a prediction that a person pays back an investment, individual level analyses show that recipients who do not repay investments in the baseline end up increasing their repayment by about 16 percent if they privately learn about such a prediction. Taken together, these results show that privately disclosing incorrect predictions to targets steers their behavior in the direction of the prediction. We interpret these findings along the lines of the identity model posited by Bénabou and Tirole (2011), which argues that when individuals observe predictions that contradict their self-perception, this may lead to a corresponding shift in self-perception and behavior.

Second, when individuals learn that investors have ignored a no-repayment prediction and invested anyway, this additionally decreases repayments. Compared to the case in which the recipient privately learns that she has been predicted not to repay, repayments decrease on average by 15 percent when the individual learns about the investor's awareness of this prediction. On an individual level, we find that this additional effect originates from participants who repay the investment in the baseline. We do not find significant differences between the cases in which only the recipient or both the investor and recipient are aware that the machine has predicted the recipient to pay back. Put differently, the investor's decisions to override the prediction that a recipient will not repay the investment reinforces the self-fulfilling prophecy. We interpret these results as evidence that overriding the no-repayment prediction provides moral "wiggle room" (Dana et al., 2007) which recipients exploit to behave more selfishly, without incurring guilt, due to the belief that investors will feel less disappointed (Battigalli & Dufwenberg, 2007, 2009).

Our results contribute to three different strands of the literature. First, we add to the literature on the interaction between humans and machines. Studies in this line of research examine the factors that determine people's reliance on algorithmic outputs, whether and

how machine learning predictions influence people's decisions, and under what circumstances their use can improve individual decision making (see e.g., Dietvorst et al., 2015; Adomavicius et al., 2018; Logg et al., 2019; Castelo et al., 2019; Yeomans et al., 2019). Cosley et al. (2003) and Adomavicius et al. (2013) find that the predictions of personalized recommendation systems may have considerable influence on users' self-reported preference ratings of products and services. There is even some evidence that personalized recommendations may create self-fulfilling prophecies by endogenously pulling consumers' willingness-to-pay in the direction of the recommendation (Adomavicius et al., 2018). Erlei et al. (2020) find that the introduction of algorithmic decision support for one party in bilateral economic bargaining may be considered as unfair by the other party, causing them to demand a better deal for themselves. De Melo et al. (2018) provide evidence that when instructing a machine to act on their behalf, people show more fairness than they would if they were to interact directly with other humans (or their machine agents). The current paper adds to this literature by showing that machine outputs may influence not only the behavior of individuals who use them as a decision-making tool, but also the behavior of targeted individuals, once they become aware of the prediction and its involvement in a decision. Our evidence accords with the notion that learning about machine predictions and their use affects participants' beliefs about (i) what kind of person they are and which social norms they are supposed to obey (Bénabou & Tirole, 2011), and (ii) how the user expects them to behave (Battigalli & Dufwenberg, 2007, 2009). That is, we complement existing work demonstrating that machine predictions can impact the construction of targeted people's beliefs.

Second, we relate to the literature that studies the consequences of informational transparency. In a field study, Ahmad et al. (2006) find that physicians are less helpful to a patient when they become aware that the patient accessed online information about her condition.[3] In an organizational setting, Inderst and Ottaviani (2012) study the interplay between dis-

[3] To the best of our knowledge this is the only field study in this literature, highlighting the difficulty of observing the effects of informational transparency in a field setting.

closing conflicts of interest and firms' strategic behavior. Similarly, but on a more individual level, there exists evidence that the disclosure of a conflict of interest in an advisor–advisee setting may lower the trustworthiness of advisors (Cain et al., 2011). In a recent paper, Inderst et al. (2019) develop and test a model of shared guilt showing that enhanced informational transparency can backfire by causing a perceived diffusion of guilt, thus crowding out prosocial behavior. Our results complement this literature by showing that providing access to machine predictions constitutes one increasingly relevant application where better access to information may unintentionally affect second order beliefs and, by extension, change behavior. Additionally, when considering the investors' point of view in our study, the change in repayment behavior due to informational transparency creates a strong incentive to follow the prediction. As a consequence, enhancing access to machine predictions may implicitly increase users' tendency to delegate real authority to the machine (Aghion & Tirole, 1997).

Third, we contribute to studies documenting the unintended downstream ramifications of integrating algorithms into social and economic systems. While many studies show how the use of machine learning applications can enhance efficiency and human welfare (see e.g., Kleinberg et al., 2018; Chalfin et al., 2016; Leo et al., 2019), there has also been a steady stream of empirical evidence on how they can facilitate and reinforce adverse outcomes. Examples include racial discrimination in the algorithmically supported recidivism decision of judges (Angwin et al., 2016), predictive policing of law enforcement units (Ensign et al., 2017), and health risk assessment of care providers (Obermeyer et al., 2019), as well as gender biases in the automated delivery of ads (Sweeney, 2013; Lambrecht & Tucker, 2019) and facial recognition tasks (Buolamwini & Gebru, 2018). This paper builds on this line of research by providing evidence that transparency concerning the nature and use of machine predictions can create moral wiggle room, thus fostering selfish behavior (Dana et al., 2007). Following our results, inaccurate machine predictions, once targets become aware of them, can be consequential even if a human decision-maker overrides and effectively renders them

mute. Simply increasing the transparency of machine predictions and their involvement in decision making processes may thus be ill-suited for alleviating organizations' accountability and transparency concerns. For instance, merely disclosing the use of machine predictions to customers may inadvertently influence their beliefs and behavior to the disadvantage of the organization.

The paper proceeds as follows. We describe our experimental design in section 2. Section 3 presents our results, while section 4 concludes.

2 Experimental design

This paper asks whether the disclosure of machine predictions and their involvement in decision making processes to affected parties (targets) influences the behavior the machine is seeking to forecast.

Several challenges arise when trying to examine this issue. First, behavioral responses are likely constrained by reputational concerns related to the repeated game nature of real life scenarios. Second, predictions by optimized machines are necessarily unique and non-random. As a consequence, it is virtually impossible to observe targets' responses to counterfactual, possibly inaccurate machine predictions, and thus identify causal relations in the field. Additionally, there are legal arguments that a purely random assignment of consequential predictions is unlawful (Parliament & Council of European Union, 2016, 2018). Third, to what extent decision-makers override or follow a machine's assessment is highly endogenous, as it depends on a variety of factors, including organizational constraints, personal preferences, and prior experience on the task.

To address these obstacles, we design a revealed-preference experimental protocol. Our design allows us to circumvent the outlined endogeneity concerns so that we can identify responses to the disclosure of predictions. The experiment comprises four subsequent stages. In stage 1, participants need to fill out a questionnaire containing 13 items on personal characteristics. Stages 2, 3, and 4 all consist of a one-shot investment game without intermediary

feedback (Berg et al., 1995).

The investment games only differ regarding investors' and recipients' access to a prediction made by a previously trained machine learning model that predicts whether recipients will repay an investment, such that the investor would be materially better off by investing. The model uses participants' questionnaire answers from stage 1 as an input. In stage 2, neither the investor nor the recipient observe the machine prediction. In stage 3, only the recipient privately learns about the prediction, while the investor remains uninformed.[4] In stage 4, the prediction becomes public information, i.e., the investor and the recipient are both aware of the prediction. We employ a strategy method in stages 3 and 4 and ask recipients to decide upon repayment for both possible cases – that is, whether the machine predicts them to repay or not. This way, we observe counterfactual decisions. Comparing recipient decisions across stages 2, 3, and 4 allows us to isolate the distinct consequences that the gradual disclosure of machine predictions may entail.

2.1 The machine learning model

The machine learning algorithm we employ is a Naive Bayesian Classifier that we previously trained, validated, and tested on a data set comprising 1397 distinct examples. We collected this data in an incentivized field study that we conducted at a large German university over three years (2016-2019).[5] Using this data, we train the algorithm to predict whether or not a person possesses reciprocal preferences. We chose a Naive Bayesian Classifier because the inner workings of this type of model are relatively intuitive and easy to understand, while at the same time offering a reasonably high predictive performance. The trained classifier we employ in our experiment uses 13 individual characteristics to make a prediction about whether an individual possesses reciprocal preferences (for more details, see the appendix). The choice of these characteristics as features is the result of comprehensive empirical testing with regard to feature selection and engineering. On a test set, the model achieves 73%

[4] Note that the investor is aware of the recipient privately learning about her prediction.

[5] We show the exact instructions of the field study in the Appendix.

accuracy.[6]

2.2 Stage 1

In the first stage of the experiment, we use a questionnaire to elicit 13 personal characteristics from participants; these characteristics serve as the input for our Naive Bayesian Classifier. At this point, we do not inform them that we will feed their answers into the trained machine learning model. In this way, we mitigate concerns participants may be motivated to give intentionally inaccurate and self-serving responses in order to "game the system".

2.3 Stage 2

Stage 2 of the experiment comprises a standard one-shot investment game (Berg et al., 1995), which serves as our individual-level baseline. The game proceeds as follows. We randomly match participants in pairs of two. First moving investors initially receive 10 MU. Investors decide whether to keep or invest the entire 10 MU. If they keep the 10 MU, the game ends. If investors decide to invest, recipients receive triple the amount, i.e., 30 MU, and need to choose an integer amount between 0 and 30 MU – that is, the sum they want to repay to the investor. We elicit recipient choices using the strategy method, i.e., we ask participants to decide when assuming the investor initially makes an investment.

2.4 Stage 3

In stage 3, participants play another investment game with a new random opponent. The only difference to the game in stage 2 is that recipients learn about the machine prediction of their behavior. We frame the prediction that a recipient possesses reciprocal preferences in terms of a strong likelihood to repay more than 10 MU to an investor (throughout this paper referred to as repayment prediction). The prediction that a recipient does not possess

[6] Note: In our experiment we define a recipient to behave reciprocally if they return an amount to the investor so that the initial investment pays off materially. According to this definition, the model correctly predicts recipients' preferences in the second stage of our experiment in 75% of the cases, i.e., there is no discrepancy between the model performance on test data and under live operation.

reciprocal preferences is presented in terms of strong likelihood to repay 10 MU or less to an investor (throughout this paper referred to as no-repayment prediction). We employ the strategy method, i.e., recipients have to make conditional decisions for both possible predictions about themselves. Recipients observe the actual prediction at the end of the experiment. Investors do not receive any information about the machine prediction beyond the fact that the recipient is aware of it when making her decision.

Employing the strategy method in this way has two advantages. First, we are able to observe recipients' incentivized responses for the actual and the counterfactual prediction of the algorithm, thus ruling out heterogeneous beliefs as a driver of the results. Second, as we do not reveal the Naive Bayes Classifier's actual prediction at this point, we ensure that participants have no reason to update their prior belief about the algorithm's predictive performance, which may affect their behavior in the subsequent stage of the experiment.

To ensure that participants understand the meaning of the prediction and enhance initial trust, we provide detailed and intuitive explanations about how the machine learning algorithm works, the data on which we trained the algorithm, and its performance on test data. In other words, we provide global explanations (Bauer et al., 2021).

2.5 Stage 4

In stage 4, participants play another investment game against a new random opponent. This game perfectly resembles the one in stage 3, except for one important difference: we inform participants that this time investors also observe the machine's prediction about the recipient before making their decision. As before, recipients have to make conditional decisions for both possible predictions about themselves.

Once participants finish stage 4, the experiment ends with a questionnaire containing several socio-demographic items. On the final screen we inform participants about the game outcomes in each stage, the machine's actual prediction about themselves, and their income.

2.6 Procedural details

In each investment game, participants always have to make choices in both the role of the investor and the recipient. The order is random. Participants do not receive intermediary feedback about the game outcome to avoid learning effects. determine the games' outcomes and payoffs in an incentive compatible way, we randomly assign investor and recipient roles and match the choices players made in the corresponding roles (conditional on the actual prediction for the recipient).

We conducted our experiment as a computerized online study, using the popular platform *Prolific* to recruit participants. The experiment is implemented using oTree (Chen et al., 2016). Overall, 156 participants took part in our experiment. On average, participants finished the experiment after 17 minutes. For every MU they possessed at the end of the experiment, we paid them 0.05 euros. On average, we paid 3.72 euros to each participant.

3 Results

In this section, we present our results in two parts. We first examine whether privately learning about the machine prediction influences recipients' repayment behavior by comparing recipient decisions in stages 2 and 3. Subsequently, our focus lies on examining how recipients responded to learning that investors' decision making involved the machine prediction. We therefore look at recipient choices in stage 4.[7]

3.1 Private disclosure of machine predictions to targets

Figure 1 (a) shows the average amount of MU that recipients repay in the case of an investment. Bar (i) portrays results for the standard investment game in stage 2 where there is no machine. Bars (ii) and (iii), respectively, depict results for the investment game in stage 3 for

[7] Note: As the analyses of investor decisions in different stages do not produce any insights that contribute to answering our main research questions and for parsimony, we refrain from reporting corresponding analyses. Instead, our focus lies on changes in recipient behaviors across distinct scenarios.

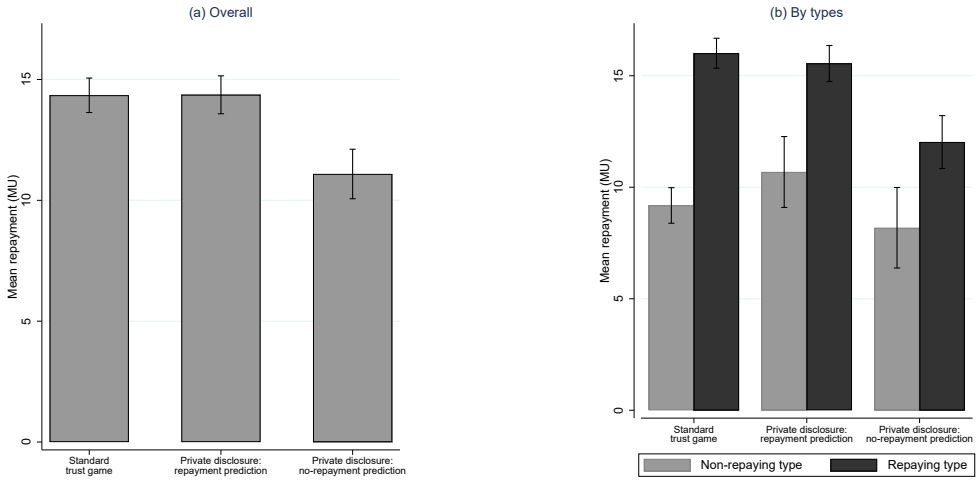

Figure 1. Mean amount of MU repaid to investor in investment games in stage 2 and 3.

the repayment (> 10 MU) and no-repayment predictions (≤ 10 MU). We show corresponding summary statistics in table 4 of the appendix.

In the standard investment game, averaged across all participants, recipients repay 14.35 MU. This amount is significantly larger than 10, so that investors are on average better off investing instead of keeping their endowment (Wilcoxon signed-rank test: $p < 0.001$). When recipients, prior to their decision, privately learn that the machine predicts them to repay more than 10 MU, they repay 14.37 MU on average. The difference to the average amount in the standard investment game is neither economically nor statistically significant (Wilcoxon signed-rank test: $p = 0.591$). However, when recipients privately learn that the machine makes a no-repayment prediction, they return 11.09 MU on average. Relative to the standard investment game baseline, this is a decline of 22.7%, which is highly significant in a Wilcoxon signed-rank test ($p < 0.001$). Notably, the average repayment amount, while significantly reduced, is is still significantly larger than 10 (Wilcoxon signed-rank test: $p = 0.037$).

Overall, aggregate level observations suggest that privately disclosing a machine prediction to targets can trigger a self-fulfilling prophecy and steer targets' behavior in the direction of the prediction. However, there seems to exist an asymmetry. We only observe such an effect in the case of a no-repayment prediction. Accordingly, the treatment the

treatment effect appears to depend on the actual prediction. To gain a better understanding of this heterogeneity, we next turn to an analysis on the individual level and look at recipient decisions conditional on the accuracy of predictions. We therefore distinguish between recipients who, in the baseline recipient decision, repay more than 10 MU (subsequently referred to as repaying types) and less or equal to 10 MU (subsequently referred to as non-repaying types).[8] According to this definition, our sample comprises 38 non-repaying and 118 repaying participants (respectively 24.36% and 75.64%). To identify individual level effects, we make use of our within subject experimental design. Figure 1 (b) depicts average repayment behavior for both types.

Dep. variable:	(1)	(2)	(3)
Repaid amount (MU)	Overall	Repaying types	Non-repaying types
Private disclosure:	0.019	-0.458	1.500*
repayment prediction	(0.437)	(0.492)	(0.904)
Private disclosure:	-3.256***	-3.983***	-1.000
no-repayment prediction	(0.437)	(0.492)	(0.904)
Constant	14.346***	16.008***	9.184***
	(0.253)	(0.287)	(0.493)
Observations	468	354	114
R^2	0.193	0.251	0.095
p	0.000	0.000	0.06

Table 1. Reported estimates result from fixed effects regression models. The dependent variable equals the recipient's repayment in the case of an investment. The explanatory variables are dummy variables that refer to different investment game scenarios. The reference category is the standard investment game. We cluster robust standard errors at the individual level and report them in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1 shows fixed-effect regression results in which the MU repayment amount serves as the dependent variable. Dummy variables representing distinct investment games serve as

[8] Note: Non-repaying and repaying participants reflect non-reciprocal and reciprocal participants, respectively. We only use this terminology in this paper.

independent variables. Repayments in the standard investment game serve as the reference category. To account for our study design, we cluster robust standard errors at the individual level and report them in parentheses. Coefficients in column (1) show results for the overall sample. Columns (2) and (3), respectively, display results for the subsample of individuals who, in the baseline, repay and do not repay an investment.

The results in column (1) corroborate our aggregate level findings. Privately disclosing to a recipient that the machine predicts her not to repay an investment leads to a significant decline in the MU repayment amount. The disclosure of a repayment prediction has a slightly positive, albeit insignificant, effect on repayment decisions.

Column (2) and (3) indicate that the aggregate level view conceals important heterogeneities. Specifically, the private disclosure of no-repayment predictions only leads to a significant decline of 24.9% in the repayment amounts for recipients who pay back more than 10 MU in the baseline (see column (2)). Recipients who repay 10 MU or less in the baseline, do not exhibit such a response. However, the estimates in column (3) indicate that these participants respond to privately learning that the machine predicts them to repay an investment by increasing the amount paid back by 1.5 MU. While this increase is only weakly significant statistically, it is of considerable economic magnitude (+16.3%).

Taken together, the regression results in table 1 suggest that privately learning what a machine predicts only affects target behavior when the prediction is inaccurate. Put differently, inaccurate predictions appear to function as a self-fulfilling prophecy: repaying participants decrease their repayment amount once becoming aware that the machine predicts them to be a non-repaying person. Conversely, non-repaying participants increase their repayment amount once becoming aware that the machine predicts them to be a repaying person.

How can we interpret this finding? The self-image model posited by Bénabou and Tirole (2011) provides one plausible explanation. According to this model, people are only dimly aware of the motives underlying their behaviors. As a result, their self-perceptions are readily influenced by the signals they receive from their environment. Specifically, when signals are in-

compatible with existing self-perceptions, individuals will revise their own self-understanding in order to accommodate these signals. Furthermore, this revised self-understanding appears to directly inform their behavior. In our experiment, when participants learn that the machine expects contrary behavior, they may interpret this as information concerning how other people with similar characteristics behave. In this way, participants may view the machine prediction as novel information about themselves, leading them to update their beliefs about what kind of person they actually are and how they are expected to behave (Krupka & Weber, 2013). In other words, the prediction may serve as a behavioral guide for participants who adjust their decisions accordingly. Therefore, the machine prediction has an "anchoring effect" (Tversky & Kahneman, 1974) that nudges the behavior of participants toward the socially "expectable" outcomes (for a literature review, see Furnham & Boo, 2011).

By contrast, accurate predictions do not invoke changed behavior, as they merely reinforce already held beliefs.

3.2 Involvement of machine predictions in decision making

Next, we examine whether recipients respond to being made aware that investors have access to machine predictions. To isolate the effect of this knowledge, we analyze differences in recipients' decisions between stages 3 and 4.

Figure 2 (a) depicts average recipient decisions conditional on the prediction and whether the prediction is only privately disclosed to the recipient or publicly disclosed to the recipient and investor alike. We report corresponding summary statistics in table 5 in the appendix.

On average, recipients repay 14.06 MU in the case they become aware that the investor observed a repayment prediction. Relative to the scenario where only recipients learn about this prediction, this is a reduction of 0.3 MU, which is neither economically (-2%) nor statistically significant (Wilcoxon signed-rank test: $p = 0.596$). Hence, on the aggregate level, recipients do not seem to respond to learning that investors had access to a prediction that the investment is materially beneficial.

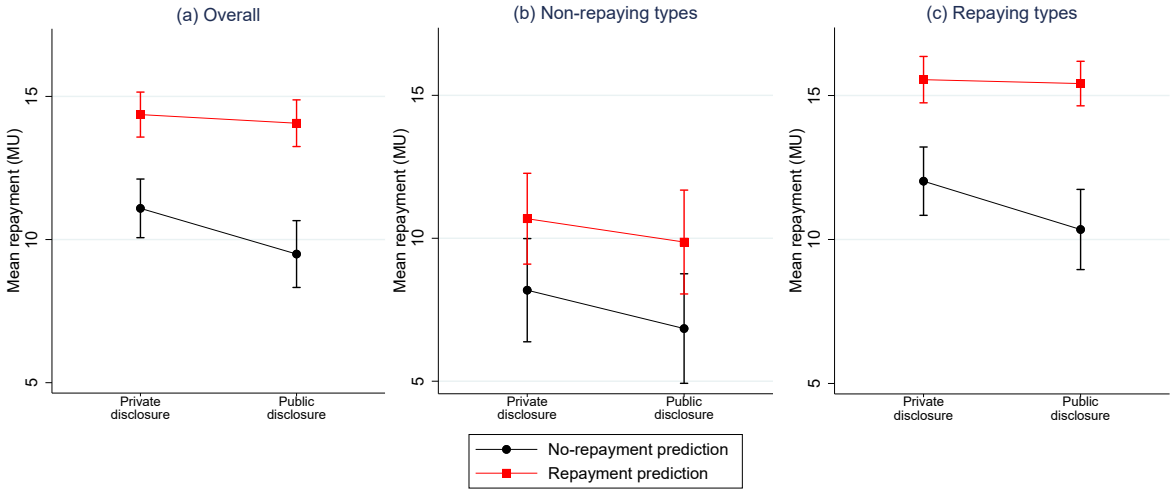

Figure 2. Mean amount of MU repaid to investor in investment games in stage 3 and 4.

By contrast, public and private disclosure of a no-repayment prediction is associated with larger divergence in repayment amounts. When recipients become aware that an investor invested despite access to a no-repayment prediction (which implicitly advised not to do so), they repay 9.49 MU on average. Compared to the case in which this prediction is private information only known to the recipient, this constitutes a reduction of 1.6 MU (-14.4%). A Wilcoxon signed-rank test confirms that this change is statistically significant ($p < 0.006$). The previously outlined self-fulfilling prophecy associated with a no-repayment prediction is therefore even more pronounced when recipients are aware that this prediction has informed the investor's decision making process.

Considering our previous individual level findings, one may naturally ask whether behavioral responses to disclosing the involvement of predictions in decision making processes depend on their accuracy (see Figure 2 (a) and (b) for an overview). To look at individual level effects, we again conduct fixed effects regression analyses, allowing us to control for participants' personal traits and inclinations. Table 2 depicts corresponding regression results. The dependent variable for all three models is the MU amount that recipients repay. We use dummy variables as regressors that refer to different investment game scenarios. The reference category is the standard investment game. We cluster robust standard errors at the

individual level and report them in parentheses. Column (1) depicts regression results for the overall sample. In columns (2) and (3) we show results for the subsamples of individuals who, according their baseline behavior, are repaying types or not.

Dep. variable:	(1)	(2)	(3)
Repaid amount (MU)	Overall	Repaying types	Non-Repaying types
Private disclosure:	0.019	-0.458	1.500*
repayment prediction	(0.376)	(0.421)	(0.795)
Private disclosure:	-3.256***	-3.983***	-1.000
no-repayment prediction	(0.493)	(0.561)	(0.967)
Additional public disclosure:	-0.301	-0.136	-0.816
repayment prediction	(0.311)	(0.314)	(0.834)
Additional public disclosure:	-1.596***	-1.678***	-1.342
no-repayment prediction	(0.504)	(0.603)	(0.909)
Constant	14.346***	16.008***	9.184***
	(0.282)	(0.319)	(0.548)
Observations	780	590	190
R^2	0.224	0.270	0.117
p	0.000	0.000	0.009

Table 2. Reported estimates result from fixed effects regression models. The dependent variable equals the recipient's repayment in case of an investment. The explanatory variables are dummy variables, referring to different investment game scenarios. The reference category is the standard investment game. We cluster robust standard errors at the individual level and report them in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

While our regression results corroborate the previously outlined aggregate level analyses (see column (1)), the aggregate level analyses once again conceal important heterogeneities. Estimates in columns (2) and (3) indicate that it is the subset of initially repaying recipients who drive the aggregate level treatment effect associated with the additional disclosure of a no-repayment prediction to the investor. Column (2) depicts that such an additional disclosure leads to a statistically significant decline in the amount of MU repaid to an investor by about 1.68 units. This additional decline is equivalent to a decrease by about 10% relative to the baseline. While the effect associated with the additional public disclosure of

a no-repayment prediction for non-repaying types is economically considerable (-14.6%), it is not statistically significant (see column (3)).[9] The additional disclosure of a repayment prediction's involvement in the investment decision does not have an effect for either type.

Overall, the regression results in table 2 depict an asymmetric effect associated with the additional disclosure of machine involvement in investor decision making.[10] Disclosing that a no-repayment prediction has been involved in the decision of an investor reinforces the self-fulfilling prophecy already associated with the private disclosure of this prediction to initially repaying types. Initially non-repaying recipients do not exhibit a statistically significant, idiosyncratic response to the disclosure of the prediction to investors.

One plausible interpretation for this finding is that recipients opportunistically use the investor decision to disregard a no-repayment prediction as an excuse to behave more selfishly without feeling guilty, because their non-repayment is not unexpected. Following the guilt aversion model, people intrinsically care about how others expect them to act. Whenever people believe they have disappointed others' expectations, they feel guilty and experience a disutility (see e.g., Battigalli & Dufwenberg, 2007, 2009). Recipients who make large repayments to investors in the standard investment game may do so because they believe that investors expect them to and thus want to avoid feelings of guilt (Khalmetski, 2016).[11] Learning that an investor invested despite observing a prediction of no-repayment allows recipients to believe opportunistically that the investor does not expect repayment. That is, the investor will feel less disappointed when not receiving repayment, which reduces the guilt felt by the recipient. As a consequence, originally reciprocal recipients can increase their

[9] Notably, a Wald test reveals that the combined effect of the private and additional public disclosure of a no-repayment prediction is statistically significant ($p < 0.025$).

[10] Note: In comparison to making decisions without a machine, investors, on average, would increase their income by 0.74 MU if they always followed the prediction. While positive, the difference is not significant in a Wilcoxon signed-rank test at any conventional level (p=0.927)

[11] Note that there is some controversial empirical evidence on the relation between prosocial behavior and guilt aversion (e.g., Dufwenberg & Gneezy, 2000; Ellingsen et al., 2010). Inderst et al. (2019) provide a refined theory supported by an experiment that is capable of reconciling much of this controversy.

overall utility by lowering repayments, as they experience less guilt–based disutility when failing to meet expectations. In other words, the information that investors deliberately disregarded the no-repayment prediction provides recipients with an excuse to behave more selfishly. Against this background, our finding relate to previous empirical work on moral wiggle rooms (see e.g., Dana et al., 2007; Andreoni & Bernheim, 2009; Grossman, 2014; Van der Weele et al., 2014).

4 Conclusion

Given the growing efforts of data privacy advocates and regulators to provide individuals affected by algorithms with a *right to explanation* about the nature and use of machine predictions, the current paper examines the consequences associated with disclosing machine predictions and their use in decision making processes. To isolate distinct causal effects associated with enhanced informational transparency, we design a parsimonious experiment specifically tailored to deal with the variety of potential confounding factors and constraints we would face in the field.

Our analyses reveal that self-fulfilling prophecies can result when machine predictions and their use are disclosed to targeted individuals. The private disclosure of an apparently incorrect prediction to recipients steers their behavior in the direction of the prediction. When initially repaying recipients additionally become aware that the investor has overridden a prediction that an investment would not pay off, they further decrease their repayment. Across all participants, disclosing to recipients that the machine predicts them not to repay and that the investor has overridden this prediction significantly reduces the repayment amount (-33.8%) such that investors are considerably worse off in this scenario. A plausible interpretation for our findings is that the disclosure of machine predictions and their use influences participants' first order beliefs about how they ought to behave, and second order beliefs about what the investor expects them to do. From this perspective, our results emphasize the role that *intelligent* prediction machines can play in the process of belief formation, and

their inherent ability to fundamentally change how people act.

Our results show that increasing the informational transparency of machine predictions for affected parties, while arguably desirable from an accountability point of view, can create unintended problems when predictions are potentially incorrect (this is not only inevitable under today's machines but it is also highly questionable whether machine predictions will ever become 100 percent accurate in the future). For example, an organization that intends to increase customers' trust and satisfaction may be ill-advised to simply reveal to customers when a process involves a machine prediction about them. That is because this measure could fundamentally change how customers behave and thus requirements for organizational strategies such as sales, customer support, and even what services are provided. These potential consequences provide an additional rationale for the installation of additional monitoring mechanisms that allows affected parties, e.g., customers, to object to algorithmic predictions about themselves they deem incorrect. Credit scoring agencies such as FICO in the US or SCHUFA in Germany already employ similar measures.

References

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? a study of anchoring effects. *Information Systems Research*, *24*(4), 956–975.

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2018). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, *29*(1), 84–102.

Aghion, P., & Tirole, J. (1997). Formal and real authority in organizations. *Journal of political economy*, *105*(1), 1–29.

Ahmad, F., Hudak, P. L., Bercovitz, K., Hollenberg, E., & Levinson, W. (2006). Are physicians ready for patients with internet-based health information? *Journal of medical internet research*, *8*(3), e22.

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, *77*(5), 1607–1636.

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

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, *97*(2), 170–176.

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, *144*(1), 1–35.

Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl (ai) n it to me–explainable ai and information systems research. *Business and Information Systems Engineering*, *2*.

Bénabou, R., & Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, *126*(2), 805–855.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122–142.

Brynjolfsson, E., & Mcafee, A. (2017). The business of artificial intelligence. *Harvard Business Review*, *7*, 3–11.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and trans-*

parency (pp. 77–91).

Cain, D. M., Loewenstein, G., & Moore, D. A. (2011). When sunlight fails to disinfect: Understanding the perverse effects of disclosing conflicts of interest. *Journal of Consumer Research*, *37*(5), 836–857.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

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

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Cosley, D., Lam, S., Albert, I., Konstan, J., & Riedl, J. (2003). Is seeing believing? now recommender interfaces affect users' opinions. In *Chi 2003* (pp. 5–10).

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80.

De Melo, C. M., Marsella, S., & Gratch, J. (2018). Social decisions and fairness change when people's interests are represented by autonomous agents. *Autonomous Agents and Multi-Agent Systems*, *32*(1), 163–187.

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.

Dufwenberg, M., & Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, *30*(2), 163–182.

Ellingsen, T., Johannesson, M., Tjøtta, S., & Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, *68*(1), 95–107.

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

Erlei, A., Nekdem, F., Meub, L., Anand, A., & Gadiraju, U. (2020). Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the aaai conference on human computation and crowdsourcing* (Vol. 8, pp. 43–52).

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, *40*(1), 35–42.

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, *60*(11), 2659–2665.

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

Horton, J. J. (2017). The effects of algorithmic labor market recommendations: Evidence from a field experiment. *Journal of Labor Economics*, *35*(2), 345–385.

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

Inderst, R., Khalmetski, K., & Ockenfels, A. (2019). Sharing guilt: How better access to information may backfire. *Management Science*, *65*(7), 3322–3336.

Inderst, R., & Ottaviani, M. (2012). Competition through commissions and kickbacks. *American Economic Review*, *102*(2), 780–809.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, *97*, 110–119.

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

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.

Kshetri, N. (2016). Big data's role in expanding access to financial services in china. *International journal of information management*, *36*(3), 297–308.

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

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

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision*

Processes, *151*, 90–103.

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

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

Parliament & Council of European Union. (2016). *Regulation (EU) 2016/679 of the european parliament and of the council.* (`https://eur-lex.europa.eu/eli/reg/2016/679/oj`)

Parliament & Council of European Union. (2018). *Regulation (EU) 2018/1807 of the european parliament and of the council.* (`https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1807`)

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

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

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

Van der Weele, J. J., Kulisa, J., Kosfeld, M., & Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American economic Journal: microeconomics*, *6*(3), 256–64.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

Supplementary Appendix

Additional material

Machine learning model

The machine learning algorithm we employ is a Naive Bayesian Classifier that we previously trained, validated, and tested on a data set comprising 1397 distinct examples. We collected this data in an incentivized field study that we conducted at a large German university over three years (2016-2019) with first-semester economics students. Most important for the experiment at hand, the field study included an incentivized one-shot prisoners' dilemma where we anonymously matched participants in pairs of two and initially endowed each one with 10 monetary units (MU). Participants could either keep the 10 MU for themselves or transfer them to their opponent. Whenever one player transferred her 10 MU, we doubled the amount so that the other player received 20 MU. Players made their choices sequentially. The second moving player received information about the first mover's choice before deciding upon the transfer herself. For each subject, we elicited both conditional choices in the role of the second mover and the unconditional choice as a first mover. In addition to the incentivized game, the field study included a broad set of survey items on students' demographics, including socio-economic background, cognitive abilities, personal traits, and other preferences. We show the exact instructions of the field study in appendix B.

According to their decisions in the sequential prisoners dilemma, we categorize participants as possessing reciprocal preferences or not (Miettinen et al., 2020). Using the categorization as labels (dependent variable) and a subset of survey responses as features (independent variables), we train a Naive Bayesian Classifier that is able to predict whether or not a person possesses reciprocal preferences. We chose a Naive Bayesian Classifier because the inner workings of this type of model are relatively intuitive and easy to understand, while at the same time reaching a reasonably high predictive performance. The trained classifier we

	Item	Scale (normalized)
1.	Big 5: Openness	(0,1)
2.	Big 5: Conscientiousness	(0,1)
3.	Big 5: Extraversion	(0,1)
4.	Big 5: Agreeableness	(0,1)
5.	Big 5: Neuroticism	(0,1)
6.	Competitiveness score	(0,1)
7.	Age in years	(0,1)
8.	Gender	Male=1, Female=0
9.	Mother possesses a college degree	Yes=1, No=0
10.	Father possesses a college degree	Yes=1, No=0
11.	Person has younger siblings	Yes=1, No=0
12.	Person has older siblings	Yes=1, No=0
13.	Person is/ was financed by parents during studies	Yes=1, No=0

Table 3. An overview of features that we use to train the Naive Bayesian Classifier. Note: we normalized the scale of numeric items for training and prediction processes.

employ in our experiment uses 13 individual characteristics to make a prediction.

Table 3 depicts the 13 distinct characteristics together with their value range. The choice of these characteristics as features is the result of comprehensive empirical testing in regards to feature selection and engineering. On a test set, the model achieves a performance of 73% accuracy. We acknowledge that we could also employ less accurate statistical methods such as a logistic or even linear regressions and that the amount of data we harness to train the model cannot be considered Big Data. The key notion, however, is that there exists a model with reasonably high predictive performance that produces a forecast intended to augment decision making. The main insights we intend to generate, namely the impact of algorithmic prediction on targets, are independent of the type of machine learning or statistical model.

Notably, in the experiment, we explicitly inform participants that we trained the machine learning system on data resulting from a game that is not exactly the same as the current investment game, but which possesses a similar design, to help investors make a payoff maximizing decision.

Investment game	Mean	Std.	Median
Standard investment game (baseline)	14.35	4.51	15
Private disclosure: repayment prediction	14.37	4.968	15
Private disclosure: no-repayment prediction	11.09	6.473	10

Table 4. Summary statistics on the amount of MU recipients repay to investors.

Δ MU repaid (Public disclosure - Private disclosure)	Mean	Std.	Median
Repayment prediction	-0.3	3.87	0
No-repayment prediction	-1.6***	6.282	0

Table 5. Summary statistics on the difference in recipient decisions between scenarios where the prediction is disclosed publicly or privately. * indicates statistical significance at the $p < 0.01$ level according to a Wilcoxon signed-rank test.

Instructions

Your task

For the following task, **a new participant of the experiment is assigned to you at random**. You and this other person make different decisions, which then result in your payout and the payout of the other person for part 2 of the experiment.

You are confronted with a task in which there are two roles (A and B). You and the other person must make a decision for both roles.

The two roles

Role A: The person in role A is given 10 points and decides first.
Role B: The person in role B has 0 points at the beginning of stage 2 and decides second. The person in role B observes the decision of the person in role A before making her/his own decision.

The person in **role A** has the following two options:
Option 1: You keep the 10 points for yourself.
Option 2: You send the 10 points to the other person. The 10 points are then tripled, i.e. the other person receives 30 points.

The person in role B is informed whether the other player has kept or sent the 10 points.

The person in **role B** has to decide how many of the points available to her/him after player A made a decision she/he wants to return to the person in role A. Player B can select any integer **between 0 and the maximum number of points at their disposal.**

Depending on the decision made by the person in role A, the person in role B can either
return 0 points (if person A chose option 1), or
return any integer between 0 and 30 points (if person A chose option 2).

Possible outcomes

Person in role A:
The total number of points of the person in role A is the sum of points person A has kept for herself/himself plus the number of points returned by B.
Example 1: Player A chooses to send 10 points, player B returns 15 points. Number of points of A = 15.
Example 2: Player A chooses to send 0 points. Hence, player B cannot return points. Number of points of A = 10.

Person in role B:
The total number of points of the person in role B is equal to the tripled number of points sent by the person in role A, minus the number of points returned to the person in role A.
Example 1: Player A chooses to send 10 points, player B returns 15 points. Number of points of B = 15
Example 2: Player A chooses to send 0 points. Hence, player B cannot return points. Number of points of B = 0

Figure 3. Instructions stage 2

Instructions summary

The two roles

The person in **role A** has the following two options:
Option 1: You keep your 10 points for yourself.
Option 2: You send the 10 points to the other person. The 10 points are then tripled, i.e. the other person receives 30 points.

The person in role B is informed whether the other person in role A has kept or sent the 10 points.

The person in **role B** has to decide how many of the points available to her/him after player A made a decision she/he wants to return to the person in role A. Player B can select any integer **between 0 and the maximum number of points at her/his disposal.**

Examples

Person in role A:
The total number of points of the person in role A is the sum of points person A has kept for herself/himself plus the number of points returned by B.
Example 1: Player A chooses to send 10 points, player B returns 15 points. Number of points of A = 15.
Example 2: Player A chooses to send 0 points. Hence, player B cannot return points. Number of points of A = 10.

Person in role B:
The total number of points of the person in role B is equal to the tripled number of points sent by the person in role A, minus the number of points returned to the person in role A.
Example 1: Player A chooses to send 10 points, player B returns 15 points. Number of points of B = 15
Example 2: Player A chooses to send 0 points. Hence, player B cannot return points. Number of points of B = 0

Your role: A

The other person takes on role B and decides second. The person in role B is informed about your decision.

You have 10 points at your disposal and the following two options to choose from:

Option 1: You keep your 10 points for yourself.
Option 2: You send your 10 points to the other person. The 10 points are tripled, i.e., the other person receives 30 points.

Note: If you decide to send your 10 points to person B, she/he can decide to return any integer between 0 and 30 points to you.

Your decision

Please make your decision and click on the "Next"-Button

Which option do you choose?

- ○ **Option 1 (keep your 10 points)**
- ○ **Option 2 (send your 10 points)**

Figure 4. Investor decision stage 2

Instructions summary

The two roles

The person in **role A** has the following two options:
Option 1: You keep your 10 points for yourself.
Option 2: You send the 10 points to the other person. The 10 points are then tripled, i.e. the other person receives 30 points.

The person in role B is informed whether the other person in role A has kept or sent the 10 points.

The person in **role B** has to decide how many of the points available to her/him after player A made a decision she/he wants to return to the person in role A. Player B can select any integer **between 0 and the maximum number of points at her/his disposal.**

Examples

Person in role A:
The total number of points of the person in role A is the sum of points person A has kept for herself/himself plus the number of points returned by B.
Example 1: Player A chooses to send 10 points, player B returns 15 points. Number of points of A = 15.
Example 2: Player A chooses to send 0 points. Hence, player B cannot return points. Number of points of A = 10.

Person in role B:
The total number of points of the person in role B is equal to the tripled number of points sent by the person in role A, minus the number of points returned to the person in role A.
Example 1: Player A chooses to send 10 points, player B returns 15 points. Number of points of B = 15
Example 2: Player A chooses to send 0 points. Hence, player B cannot return points. Number of points of B = 0

Your role: B

The other person in role A makes the decision first.
You must decide what you want to do, for the case that the person in role A sends you the 10 points so that you have 30 points at your disposal.

Your decision

Please make your decision, assuming the person in role A sent you the 10 points before, i.e., you have 30 points at your disposal. Please make your decision now and then press the "Next" button.

How many points would you like to return to person A,
if she/he has sent you 10 points, i.e., you have 30 at your disposal?
(Please enter any number between 0,1,2,3,4,5,...,30):

Figure 5. Recipient decision stage 2

Your task

For the following task, **a new participant of the experiment is assigned to you at random**.
The decision situation in part 3 is similar to the previous one in part 2, i.e., you are again confronted with two decisions in which you respectively take on the roles A and B.
However, there is one important difference:

The person in role B receives a prediction, generated by a Machine Learning System, about herself/himself saying whether or not she/he is classified as being likely to return 10 or more points, when player A decides to send the 10 points. The prediction about a person is generated using this person's questionnaire answers.
The Machine Learning System asks: 'given a person's characteristics, is this person in role B most likely to return 10 or more points, when they receive 10 points from the person in role A?'

In a test, the trained Machine Learning System correctly predicts participants' tendency to behave reciprocally in more than 70% of the cases.

Important: The person in role A does NOT observe this prediction about the other person in role B. Hence, the person in role A cannot rely on the prediction to make her/his decision.

The prediction about a person is generated using this person's questionnaire answers. It is completely independent of the person's decisions in previous parts of the experiment.

The Machine Learning System was trained and tested on 1039 distinct observations generated by participants in previous experiments. The Machine Learning System learned to make a prediction about people's tendency to return points based on observations from a task that has the same general structure, but is slightly different from the task in stage 3.

Below you can find additional information about the structure of the system. We use the Bernoulli Naive Bayes Machine Learning model.

Different scenarios

You and the other person who is matched with you will have to make a decision for both roles A and B.

Figure 6. Instructions stage 3

Additional Information about the Machine Learning System

The Naive Bayes Model is one of the simplest, yet one of the most powerful machine learning algorithms for classification. In the context of the experiment, classification refers to correctly predicting whether a person in role B will return 10 or more points or not, given player A sends the 10 points.

Note: The Bayes Model used in this experiment was trained using data from decision situations that had a slightly different structure than the one you encounter here. In particular, the decision scenario the machine was trained on had the following structure.

There were two players A and B that were matched and initially endowed with 10 points. Both players had to choose individually between (i) keep the 10 points for themselves, or (ii) send the 10 points to the other person. If the 10 points were send, they were doubled, i.e. the other player received 20 points. Player A had to make the choice first and without knowing what player B will do. Player B made the choice second and was informed about the choice of player A before making the own decision.
The machine was trained to predict how the player in role B would behave, in case player A sends 10 points. In other words, the machine learned to predict whether or not player B behaves reciprocally and sends points back to player A, in case player A sends 10 points.

The Naive Bayes Model is based on Bayes' theorem. Bayes' theorem describes the probability that an event occurs, given prior information about conditions that might be related to the event. It serves as a way to figure out conditional probability. In this experiment, the Naive Bayes Model uses participants survey answers to compute the probability that they are sending back points in case you send them 10 points. In other words, the Naive Bayes Model learned the underlying relationship between people's characteristics and their behavior.

Three examples of real world applications:

Online customer behavior
Naive Bayes classifiers are very effective in predicting people's activity and behavior on the internet. Generated prediction are often very accurate and frequently used in Product Recommendation, Advertisement, and many more.

Spam Filters
Naive Bayes classifiers are a popular and effective technique for e-mail filtering. The Naive Bayes Model works by correlating the use of words with spam and non-spam e-mails and then, using Bayes' theorem, calculate a probability that an email is or is not spam.

Medical Diagnosis
One of the main advantages of the Naive Bayes approach which is crucial for medical diagnosis is that all the available information is used to make a prediction. When dealing with medical data, the Naive Bayes Model takes into account evidence from many attributes to make the final prediction. Additionally, it provides transparent explanations of its decisions, which is why it is considered one of the most useful classifiers to support physicians' decisions.

Figure 7. Additional information algorithm

Summary Machine Learning System

The prediction

The Machine Learning System asks:
Given player B's characteristics, will she/he most likely return 10 or more points to player A, when player A sends the 10 points?

Performance

In a test, the trained Machine Learning System correctly predicts participants' propensity to return points in more than 70% of the cases.

Training process

The Machine Learning System is a Naive Bayes Model that learned the underlying relationship between people's characteristics and their behavior as player B. The Naive Bayes Model is one of the simplest, yet one of the most powerful machine learning algorithms to make predictions. We trained and tested the machine on 1039 distinct observations generated by participants in previous experiments.

Your role: B

Note: The other person in role A does NOT observe the Machine Learning System's prediction about yourself.
Therefore, the other person in role A cannot rely on the prediction to make her/his decision.

The other person in role A makes the decision first.
You must decide what you want to do, for the case that the person in role A sends you the 10 points so that you have 30 points at your disposal.

Your decisions

Please make your decision, assuming the person in role A sent you the 10 points before, i.e., you have 30 points at your disposal.

You have to decide how many points to return to the player in role A for both possible predictions the Machine Learning System can make about you.

You will learn about the actual prediction at the end of the experiment.

Figure 8. Recipient decision stage 3

Possible scenario 1:
Given your questionnaire answers, the Machine Learning System predicts that:
you will most likely NOT return 10 or more points.

In this scenario,
how many points would you like to send back to person A,
if she/he has sent you 10 points before, i.e., you have 30 points at your disposal?
(Please enter any number between 0,1,2,3,4,5,...,30):

Figure 9. Recipient decision stage 3

Possible scenario 2:
Given your questionnaire answers, the Machine Learning System predicts that you will:
you will most likely return 10 or more points.

In this scenario,
how many points would you like to send back to person A,
if she/he has sent you 10 points before, i.e., you have 30 points at your disposal?
(Please enter any number between 0,1,2,3,4,5,...,30):

Figure 10. Recipient decision stage 3

Your task

For the following task, **a new participant of the experiment is assigned to you at random**.
The decision situation in part 4 is similar to the previous one in part 3, i.e. you are again confronted with two decisions in which you respectively take on the roles A and B and there is a Machine Learning System that generates a prediction.
The Machine Learning System, and the prediction are exactly the same as in the previous stage.

The only difference to stage 3 is the following:

Now the other person in role A learns about the machine-generated prediction about whether or not the person in role B is most likely to return 10 or more points, when player A decides to send the 10 points.

This means that now the person in role A can, but does not have to, rely on the Machine Learning System's prediction to make her/his decision.

Different scenarios

You and the other person who is matched with you will have to make a decision for both roles A and B.

Figure 11. Instructions stage 4

Summary Machine Learning System

The prediction

The Machine Learning System asks:
Given player B's characteristics, will she/he most likely return 10 or more points to player A, when player A sends the 10 points?

Performance

In a test, the trained Machine Learning System correctly predicts participants' propensity to return points in more than 70% of the cases.

Training process

The Machine Learning System is a Naive Bayes Model that learned the underlying relationship between people's characteristics and their behavior as player B. The Naive Bayes Model is one of the simplest, yet one of the most powerful machine learning algorithms to make predictions. We trained and tested the machine on 1039 distinct observations generated by participants in previous experiments.

Your role: A

The other person takes on role B and decides second.
The person in role B is informed about your decision and that you saw the Machine Learning System's prediction about her/his own propensity to return points.

You see the prediction about the other person and you can, but you do not have to, rely on the prediction to make your decision.

You have 10 points at your disposal and the following two options to choose from:

Option 1: You keep your 10 points for yourself.
Option 2: You give your 10 points to the other person. The 10 points are tripled, i.e. the other person receives 30 points.

Note: If you decide to send your 10 points to person B, she/he can decide to return any integer between 0 and 30 points to you.

The prediction about the other person in role B

Given the other person's questionnaire answers, the Machine Learning System predicts:

The other person in role B will most likely return 10 or more points to you, if you send your 10 points

Your decision

Please make your decision and click on the "Next"-Button

Which option do you choose?

- **Option 1 (keep your 10 points)**
- **Option 2 (send your 10 points)**

Figure 12. Investor decision stage 4

Summary Machine Learning System

The prediction

The Machine Learning System asks:
Given player B's characteristics, will she/he most likely return 10 or more points to player A, when player A sends the 10 points?

Performance

In a test, the trained Machine Learning System correctly predicts participants' propensity to return points in more than 70% of the cases.

Training process

The Machine Learning System is a Naive Bayes Model that learned the underlying relationship between people's characteristics and their behavior as player B. The Naive Bayes Model is one of the simplest, yet one of the most powerful machine learning algorithms to make predictions. We trained and tested the machine on 1039 distinct observations generated by participants in previous experiments.

Your role: B

Note: The other person in role A observes the Machine Learning System's prediction about yourself before making her/his decision.
Therefore, the other person in role A can, but does not have to, rely on the prediction to make her/his decision.

The other person in role A makes the decision first.
You must decide what you want to do, for the case that the person in role A sends you the 10 points so that you have 30 points at your disposal.

Your decisions

Please make your decision, assuming the person in role A has seen the prediction and sent you the 10 points before, i.e., you have 30 points at your disposal.

You have to decide how many points to return to the player in role A for both possible predictions about yourself that the other person could have seen.

You will learn about the actual prediction at the end of the experiment.

Figure 13. Recipient decision stage 4

Possible scenario 1:
Given your questionnaire answers,
the Machine Learning System informed the other player in role A that:

you will most likely NOT return 10 or more points, if she/he sends you 10 points.

In this scenario,
how many points would you like to send back to person A,
if she/he has sent you 10 points before, i.e., you have 30 points at your disposal?
(Please enter any number between 0,1,2,3,4,5,...,30):

Figure 14. Recipient decision stage 4

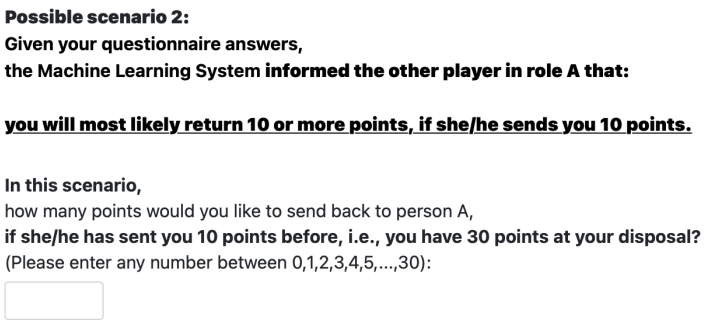

Figure 15. Recipient decision stage 4

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 312	Can Gao Ian Martin	Volatility, Valuation Ratios, and Bubbles: An Empirical Measure of Market Sentiment
No. 311	Wenhui Li, Christian Wilde	Separating the Effects of Beliefs and Attitudes on Pricing under Ambiguity
No. 310	Carmelo Latino, Loriana Pelizzon, Aleksandra Rzeźnik	The Power of ESG Ratings on Stock Markets
No. 309	Tabea Bucher-Koenen, Andreas Hackethal, Johannes Koenen, Christine Laudenbach	Gender Differences in Financial Advice
No. 308	Thomas Pauls	The Impact of Temporal Framing on the Marginal Propensity to Consume
No. 307	Ester Faia, Andreas Fuster, Vincenzo Pezone, Basit Zafar	Biases in Information Selection and Processing: Survey Evidence from the Pandemic
No. 306	Aljoscha Janssen, Johannes Kasinger	Obfuscation and Rational Inattention in Digitalized Markets
No. 305	Sabine Bernard, Benjamin Loos, Martin Weber	The Disposition Effect in Boom and Bust Markets
No. 304	Monica Billio, Andrew W. Lo, Loriana Pelizzon, Mila Getmansky Sherman, Abalfazl Zareei	Global Realignment in Financial Market Dynamics: Evidence from ETF Networks
No. 303	Ankit Kalda, Benjamin Loos, Alessandro Previtero, Andreas Hackethal	Smart (Phone) Investing? A Within Investor-Time Analysis of New Technologies and Trading Behavior
No. 302	Tim A. Kroencke, Maik Schmeling, Andreas Schrimpf	The FOMC Risk Shift
No. 301	Di Bu, Tobin Hanspal, Yin Liao, Yong Liu	Risk Taking, Preferences, and Beliefs: Evidence from Wuhan
No. 300	Dennis Gram, Pantelis Karapanagiotis, Jan Krzyzanowski, Marius Liebald, Uwe Walz	An Extensible Model for Historical Financial Data with an Application to German Company and Stock Market Data
No. 299	Ferdinand A. von Siemen	Motivated Beliefs and the Elderly's Compliance with COVID-19 Measures