Kevin Bauer | Moritz von Zahn | Oliver Hinz

Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Cognitive Processes

SAFE Working Paper No. 315

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

Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Cognitive Processes

Kevin Bauer* Moritz von Zahn† Oliver Hinz†

June 16, 2021

Abstract

This paper explores the interplay of feature-based explainable AI (XAI) techniques, information processing, and human beliefs. Using a novel experimental protocol, we study the impact of providing users with explanations about how an AI system weighs inputted information to produce individual predictions (LIME) on users' weighting of information and beliefs about the task-relevance of information. On the one hand, we find that feature-based explanations cause users to alter their mental weighting of available information according to observed explanations. On the other hand, explanations lead to asymmetric belief adjustments that we interpret as a manifestation of the confirmation bias. Trust in the prediction accuracy plays an important moderating role for XAI-enabled belief adjustments. Our results show that feature-based XAI does not only superficially influence decisions but really change internal cognitive processes, bearing the potential to manipulate human beliefs and reinforce stereotypes. Hence, the current regulatory efforts that aim at enhancing algorithmic transparency may benefit from going hand in hand with measures ensuring the exclusion of sensitive personal information in XAI systems. Overall, our findings put assertions that XAI is the silver bullet solving all of AI systems' (black box) problems into perspective.

Keywords: XAI, explainable machine learning, Information Processing, Belief updating, algorithmic transparency

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

†Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany, E-Mail: vzahn@wiwi.uni-frankfurt.de, ohinz@wiwi.uni-frankfurt.de

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

1 Introduction

Organizations increasingly harness artificial intelligence (AI) systems to augment the decision making of their employees, mainly by producing predictions that aim at mitigating informational asymmetries. For instance, AI systems support hiring decisions by generating candidate performance predictions (Hoffman et al. 2018), medical diagnosing by providing predictions about afflictions (Esteva et al. 2019), bail decisions by providing recidivism predictions (Kleinberg et al. 2018), financial investments by generating profitability forecasts (De Spiegeleer et al. 2018).

While state-of-the-art AI methods achieve unprecedented prediction accuracy (see e.g., Jordan and Mitchell 2015), the high predictive performance usually comes at the expense of understanding how the information going in, i.e., input features, relate to the information going out, i.e., the prediction (Páez 2019, Arrieta et al. 2020). Put differently, the inner workings of most contemporary AI methods, e.g., Deep Neural Networks or Random Forests, are unintelligible to human users. This "black box" nature can have considerable downsides (Bauer et al. 2021), including the impairment of trust in and reliance on machine outputs by users (Wang and Benbasat 2007, Kizilcec 2016), the reduced likelihood to detect incorrect behaviors and biases (Roselli et al. 2019), and the hindrance of knowledge-transfers from machines to humans (Rosenfeld and Richardson 2019, Vilone and Longo 2020).

The shortcomings associated with black box machines have sparked public and regulatory interest. Policy makers increasingly demand that "meaningful explanations of the logic involved" [1] need to accompany AI systems' predictions as a means to alleviate black box problems. Examples include Europe's (and UK's) GDPR regualation or its recent proposal for an Artificial Intelligence Act, the Equal Credit Opportunity Act in the US, or the Digital Republic Act in France. Hence, the importance of employing explainability measures in organizations steadily grows. Against this background, it is not surprising that researchers' and practitioners' increasingly focus on the development of explainable AI (XAI) methods that elucidate AI systems' inner logic (Ji-Ye Mao 2000, Adadi and Berrada 2018, Lakkaraju et al. 2019, Bhatt et al. 2020). A large part of contemporary XAI revolves around feature-based, local explainability. That is, "machine explanations" about how an AI system weighs given features and arrives at a specific prediction (see e. g., LIME (Ribeiro et al. 2016a) or SHAP (Lundberg and Lee 2017)).

Despite the importance to provide machine explanations on how AI systems produce specific predictions, researchers have only recently started to explore the prerequisites and consequences of human-XAI interaction (see e.g., Yang et al. 2020, Liao et al. 2020, Lakkaraju and Bastani

[1] See EU (2021)

2020). There are many open questions, especially concerning the interplay of providing machine explanations, human behavior, and human cognitive processes. The paper at hand contributes to this nascent, but important, literature. We examine the impact of feature-based explanations on human information processing and human beliefs, using LIME (Ribeiro et al. 2016a,b). LIME is one of today's most popular XAI techniques that explains why AI systems produce individual predictions in an intuitive, graphical way. Specifically, LIME typically uses colored bar charts to represent the weights given to input features for individual predictions. Bars' colors and lengths signal to users whether associated input features are evidence in support of or against the produced prediction (Ribeiro et al. 2016a). Notably, while LIME explanations, as well as other feature-based explanations, and coefficients from linear or logit regression are somewhat similar because both depict the relation between distinct features (independent variables) and the label (dependent variable), they differ in one key aspect: the importance of a specific feature (or independent variable) and its influence on the label (dependent variable) can vary for every single observation. That is, LIME tries to explain the relation between features and labels for individual observations (locally), while linear or logit regressions explain the aggregate relation between features and labels for the overall data (globally). By means of LIME explanations, we study whether feature-based explainability techniques lead users to (i) change their weighting of available information when making a decision, (ii) adjust their preexisting beliefs about the relation between inputted information and the prediction target, and (iii) whether users' trust in the AI-sytem's prediction accuracy moderate belief adjustment processes.

As a tangible example of the settings we have in mind, consider a bank's loan officer who decides upon lending money to an applicant. To inform her decision, she observes personal characteristics of the applicant (e.g., age, credit history, occupation, ...) that she considers differently related to the applicant's repayment probability and thus relevant for the approval decision. Additionally, she observes a prediction about the applicant's creditworthiness from an AI. Due to novel regulations, the bank decided to employ a feature-based explainability tool. This tool explains how applicants' personal characteristics contribute to individual predictions about the creditworthiness. The officer, for example, in addition to the applicants characteristics, might now observe that the AI system predicts the current applicant not to repay a loan mainly due to his age and gender. Regarding this example, our study intends to answer the following three questions: (i) Does the provision of machine explanations change the loan officer's weighting of observed borrower characteristics when making her decision? (ii) Does the provision of explanations change her preexisting beliefs about the relation between observed characteristics and repayment behavior? (iii) What is the role of the officer's trust in

the prediction accuracy of the AI-system when it comes to belief adjustments?

Considerable challenges arise when trying to answer these questions. First, identifying how machine explanations affect users' weighting of available information is extremely difficult because it is generally an unobserved cognitive process influenced by a multitude of external cues that one needs to control. Second, people's (preexisting) beliefs depend on abundant unobserved factors (e.g., personal taste, experience, or domain knowledge) that we need to control for when examining the influence of machine explanations on beliefs. Third, whether people in organizations are willing to interact with an (explainable) AI in the first place, let alone rely on it, is highly endogenous and depends on factors such as the organizational culture, degree of digitization, and technological literacy. Overall, these challenges are extremely difficult, if not impossible, to overcome in a natural field setting. To identify the causal effect of explanations given by an AI on human information processing and belief structures, we design a novel experimental protocol that we carry out as an online study. The design closely mirrors the fundamental structure of many AI-augmented decision making processes in organizations while at the same time providing us the necessary control over potential confounds to answer the research questions at hand.

Our experiment comprises five different stages. In each stage, participants engage in several incentivized investment games (Berg et al. 1995). Participants act as investors and decide whether to invest material resources with different borrowers, who may or may not repay the investment. Before making a decision, investors observe the personal characteristics of a borrower. In stages I and II, we respectively elicit participants' prior investment behavior and beliefs about the most decision-relevant borrower characteristics. Stage III serves as our treatment manipulation, where treatment participants make a series of investment decisions with the additional aid of an explainable AI that shows the individual contribution of borrower characteristics to the overall prediction in the form of LIME values (Ribeiro et al. 2016a). Baseline participants interact with an opaque AI. We elicit participants' posterior beliefs and investment behavior in stages IV and V, respectively. By comparing how participants make decisions before, during, and after interacting with the explainable or opaque AI, we can isolate how machine explanations impact their belief structures.

The paper proceeds as follows. Section 2 summarizes the existing literature we contribute to. In section 3, we outline our empirical strategy. We present our results in section 4. Section 5 and 6, respectively, discusses our findings and concludes.

2 Related literature

Our work addresses open research questions from three different strands of the literature. The first and most closely related line of work studies the interplay between explainable AI techniques and user behavior (see Doshi-Velez and Kim 2017, Vilone and Longo 2020, for a review). The foundation of this literature has already been laid about two decades ago with early explorations of how explanations about the functioning and purpose of recommender systems affect their use and the decision making of users (Dhaliwal and Benbasat 1996, Gregor and Benbasat 1999, Wang and Benbasat 2007). With the development of modern XAI techniques, research on the consequences of employing explainability techniques has seen a considerable resurgence (Vilone and Longo 2020). Much of this current work considers the impact of explanations on user trust and reliance on AI systems. Yin et al. (2019) find that revealing test data accuracy affects user trust in AI systems, with the effect size being smaller after observing the live accuracy of the system. In the context of clinical decision support systems, Bussone et al. (2015) find that overly detailed explanations about how the system arrives at a certain output can enhance trust but may also create overreliance on the recommendations. In contrast, short or absent explanations appear to foster overreliance but decrease trust. In an experimental study, Erlei et al. (2020) find some evidence indicating that global explainability about the model and its performance may have negative effects on user trust. Relatedly, Zhou et al. (2017) find that user trust in the AI decreases significantly once they become aware of the uncertainty associated with predictions. Other studies focus on the consequences of XAI techniques for decision making. There is evidence that explanation complexity increases users' time to act upon machine outputs (Narayanan et al. 2018). Yang et al. (2020) find that the additional provision of visual example-based explanations about why a machine learning model makes a specific classification improves the performance of the human-machine collaboration. Poursabzi-Sangdeh et al. (2021) find that feature-based explanations can have negative effects on detecting erroneous predictions, and thus on performance, due to information overload. The aforementioned studies make important contributions towards understanding the consequences of XAI techniques for decision making. However, there are many open questions, especially when it comes to how XAI techniques affect users' fundamental cognitive processes, i.e., the psychological underpinnings of observed effects on decision making. Our study contributes to filling this gap by outlining how feature-based XAI influences users' cognitive weighting of available information that are relevant to the decision. This way, we are able to provide better insights why feature-based XAI influences decision making. We even go one step further and examine whether, and if so how, feature-based XAI techniques endogenously influence human beliefs about the relationship between

input features and the decision problem. Exploring the existence of such effects can improve our understanding about how XAI techniques can serve as a tool to transfer knowledge domain-knowledge from machines to humans. To answer our research questions, we make use of an incentivized experiment that allows us to isolate causal effects from feature-based XAI on human cognitive processes. To the best of our knowledge, we are the first to address these open gaps.

The second stream of literature that we complement, studies how people respond to and make use of algorithmic recommendations. There has been a steady stream of research documenting that humans tend to discount the advice by machines relative to the advice by fellow humans, even when advice-takers are aware that machine advice is more accurate (Meehl 1954, Dawes et al. 1989, Grove and Meehl 1996, Grove and Lloyd 2006, Önkal et al. 2009, Dietvorst et al. 2015). The predilection to discount machine advice disproportionately has recently been coined as algorithm aversion (Dietvorst et al. 2015). This line of work broadly suggests that people possess an inherent distrust towards algorithmic outputs (see Burton et al. 2020, for a review). Several factors moderate the occurrence of algorithm aversion. Examples include the perceived subjectivity of the task (Yeomans et al. 2019, Castelo et al. 2019), seeing the algorithmic output being incorrect (Dietvorst et al. 2015, Prahl and Van Swol 2017), being able to modify predictions (Dietvorst et al. 2018), and the degree actual and expected predictive performance diverge (Bhattacherjee and Premkumar 2004, Jussupow et al. 2020). More recently, studies by Logg et al. (2019), Gunaratne et al. (2018), Prahl and Van Swol (2017) have found that there are also domains and scenarios in which humans prefer algorithmic over human advice (algorithm appreciation coined by Logg et al. (2019)). This result suggests that algorithm aversion is neither a universal nor a straightforward phenomenon. Despite recent advances in this field, it remains open whether, and if so how, XAI techniques affect people's aversion of or preference for algorithmically produced recommendation. XAI may introduce another layer of complexity that influences people's attitude towards algorithmic advice. Our paper contributes to this literature by showing how the introduction of feature-based explanations affects people's reliance on predictions. Gaining a better understanding of the interplay between XAI techniques and algorithm aversion is particularly important from a practitioners perspective as there is a growing number of regulatory requirements stipulating that the output of AI systems, in a multitude of domains, needs to be human-interpretable (see e.g., GDPR 2016, EU 2021).

Finally, we contribute to previous research that explores the underlying mechanisms of information processing and belief updating. Here a common theoretical foundation builds upon Bayes rule as a rational benchmark to incorporate new information into beliefs based on a

weighting of new signals and prior beliefs (see e.g., Slovic and Lichtenstein 1971). However, research has shown systematic deviations from Bayes' rule (Epstein et al. 2010, Rabin 2013). Reasons include primacy and recency effects (Hogarth and Einhorn 1992), base rate neglects (Kahneman and Tversky 1973), egocentric underweighting of new information (see e.g, Rabin and Schrag 1999, Yaniv 2004), and a general tendency to both discount information conflicting with prior beliefs and readily internalize information in line with prior beliefs (confirmation bias, Edwards and Smith 1996, Nickerson 1998, Ditto and Lopez 1992, Rabin and Schrag 1999). Important moderating factors include the perceived expertise of the advisor (see e.g., Pilditch et al. 2020) and the distance between the advice and preexisting beliefs (see e.g., Yaniv 2004). While there is ample evidence on how and why people deviate from rational belief updating in human-human collaborative settings, it is unclear to what extent observed patterns apply in human-XAI settings. For example, does trust in the predictive performance of an AI system moderate belief updating evoked by observed machine explanations? A thorough understanding of belief updating through XAI methods is a necessary prerequisite to harness AI systems' broader potential of "machine teaching", the notion that humans learn from AI systems (Abdel-Karim et al. 2020).

In this paper, we aim to contribute to closing the aforementioned research gaps by exploring the causal relationships between explainable AI, information processing, and belief updating. This includes a deeper analysis of moderating factors such as trust in the predictive performance and the prior belief structure of users. As a result, we complement the broader discussion about how AI affects human decision making (e. g., Jussupow et al. 2021) and provide important insights frequently requested by leading IS scholars (e. g., Ågerfalk 2020). Our work is particularly relevant in light of recent studies underlining the potential of explainable AI to mislead users (Lakkaraju and Bastani 2020).

3 Empirical strategy

We aim to examine the interplay between XAI, human behavior, and human beliefs. More specifically, this paper intends to answer three questions: (i) Does the provision of machine explanations change users' weighting of available information when making decisions? (ii) Does the provision of machine explanations affect users preexisting beliefs about the relation between observed information and a target variable? (iii) Does trust in the explainable AI's prediction accuracy moderate the adjustment of beliefs?

There are considerable challenges when trying to answer this question. First, studying how people weigh information is extremely challenging because it is a mental process often occurring

to a considerable extent subconsciously. Additionally, people generally process a multitude of external, and even internal cues (e.g., intuition or gut feeling), that are difficult to identify, not to mention objectively measure. Second, measuring (changes in) beliefs is inherently difficult because their initial adoption and update are unobserved cognitive processes that depend on a variety of factors including the socio-cultural environment (see e.g., Kruglanski 1996), prior experience (see e.g., Rabin and Schrag 1999), and strategic consideration (see e.g., Zimmermann 2020). The identification of effects associated with providing machine explanations on human beliefs requires tight control over these covariates. Third, whether organizations employ (explainable) AI to augment human decision making is highly endogenous, depending on factors such as the organizational culture, type of tasks, and the degree of organizational digitization. These factors in addition to other organizational-strategic concerns determine employees' readiness to engage with, rely on, and learn from the technology (see e.g., Venkatesh and Davis 2000, Venkatesh et al. 2012). Answering our research question requires tight control over these factors.

In the paper at hand, we address these challenges by designing an incentivized, revealed-beliefs experiment that we implement as an online study. We tailor our experiment to mirror the fundamental structure of many AI-in-the-loop decision making processes in organizations, while at the same time providing us with the required control over the aforementioned covariates that we could not obtain in a field setting. With this approach, we are in line with previous IS research that has successfully built upon experiments in controlled lab environments (see e. g., Jiang and Benbasat 2004, 2007, Adam et al. 2015).

3.1 Experimental design

The experiment comprises 5 subsequent stages (see Figure 1 for an overview). In each stage, participants repeatedly engage in a modified version of the one-shot investment game (Berg et al. 1995) that possesses the following structure. An investor and a borrower possess an initial endowment of 10 monetary units (MU). The investor initially observes up to ten of the borrower's characteristics and decides whether to invest her 10 MU with the borrower or keep the 10 MU for herself. If the investor keeps her endowment, both the investor and borrower receive a payoff of 10 MU. If she invests her endowment, the borrower receives 20 MU and has to decide whether or not to repay the investor by giving up 10 MU. In case of repayment, the investor receives 20 MU so that the initial investment pays off; otherwise the investor ends up with 0 MU while the borrower earns 30 MU (see Figure 2). This investment game mimics the fundamental structure of many sequential, strategic decisions under uncertainty

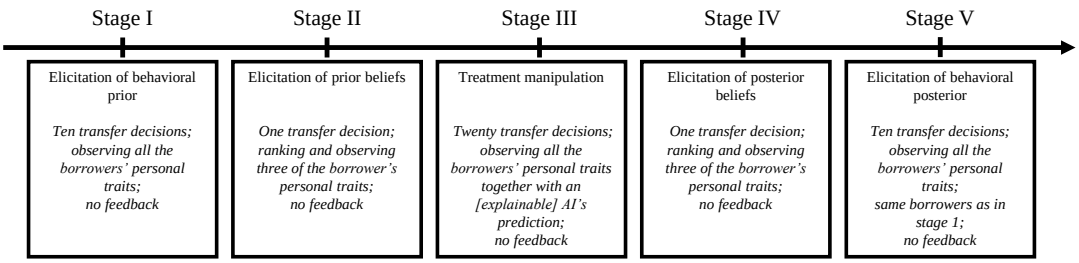

Figure 1: Sequence of the experiment

Notes: Sequence and overview of the 5 different stages in the experiment. We denote the treatment variation in stage III via the expression [explained].

(e.g., lending decisions, market transactions, and hiring decisions) (Fehr and Fischbacher 2003) while at the same time providing a level of abstraction that mitigates concerns about investors' prior task-related knowledge and stereotypes. At the end of the experiment, we pay investors and borrowers according to game outcomes, i.e., the experiment is incentivized allowing us to measure revealed preference which is superior to purely self-reported answers (Camerer and Hogarth 1999).

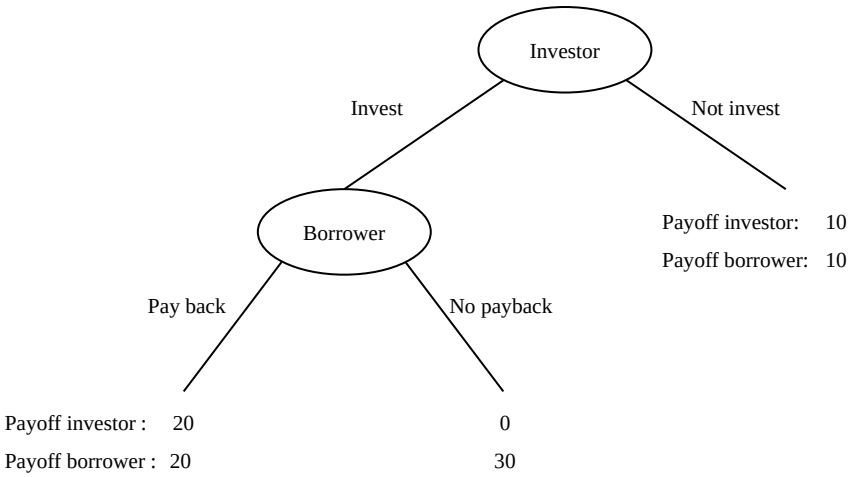

Figure 2: Trust game structure

Notes: Structure of the modified trust game employed as the main workhorse throughout the experiment.

In a nutshell, our experiment works as follows. There are five subsequent stages, with every

single stage being individually incentivized. In stages I and II, we respectively elicit participants' prior investment behavior and developed beliefs about the most decision-relevant characteristics by letting them make several investment decisions without intermediary feedback. In stage III, investors make another series of decisions with the additional aid of an AI that provides predictions about the borrowers' repayment behavior and, depending on the experimental condition, comes with or without explanations about how the observed characteristics relate to the prediction. Stages IV and V respectively mirror stages II and I, allowing us to elicit investors' posterior behavior and beliefs. We show the developed interfaces in Appendix B. To prevent the development of expertise, idiosyncratic investment strategies, and path dependencies based on the consequences of investment decisions that might confound our results, we do not provide intermediary feedback. Comparing changes in investor beliefs after they interacted with the AI across the baseline and treatment condition allows us to isolate the effects of machine explanations on user beliefs.

3.2 Borrowers, the AI, and LIME explanations

Participants in our online study always take on the role of the investor. Borrowers are subjects from a previous incentivized field study where we elicited repayment decisions using the strategy method, i.e., participants had to decide upon repayment under the assumption that their opponent initially invests. More specifically, the field study comprises a variation of an incentivized one-shot investment game and a broad set of survey items on participants' demographics, socio-economic background, cognitive abilities, and other personality traits. Overall, we collected more than 2,500 individual observations over three years (2016-2019). For our online study, we use 1,104 distinct observations of this data set.[2]

In preparation for the online study, we randomly split the 1104 observations into two representative subsets: a training set (n=1054) and a player set (n=50).[3] We use the training set to build a Gradient Boosted Random Forest (GBRF) that uses ten socio-demographic borrower characteristics to predict whether or not a person will repay an investment (see Table 4 in the appendix).[4] Investors in our online study always observe these ten borrower characteristics before making their decision. The rationale for choosing these very ten features is twofold. On the one side, we wanted to include borrower characteristics accessible to investors, i.e., that they could intuitively relate to a borrower's repayment behavior. Additionally, we worked on building

[2]After careful cleaning and preprocessing of the overall data set, we are left with 1,104 observations that we are confident to use for the online study. For more details on the field study, see the Appendix A.

[3]Note: a Kolmogorov-Smirnov test cannot reject the hypothesis that both sets stem from the same underlying population $p = 0.781$

[4]The characteristics are: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism, Competitiveness, Patience, Gender, having younger siblings, and having older siblings.

a model that has reasonably high predictive performance, so that the choice of these features, at least partially, result from the comprehensive empirical tuning of the GBRF.[5] We render the "black box" GBRF model explainable, using feature-based explanations provided by the Python library *InterpretML* (Nori et al. 2019), an open-source package that incorporates state-of-the-art machine learning explainability techniques. Specifically, we generate local feature-based explanations about why the AI system produces individual predictions for the player set using the model-agnostic surrogate technique LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al. 2016a). LIME is one of the most popular and widely used explainability techniques as of today (see e.g., Feng and Boyd-Graber 2019, Bhatt et al. 2020). LIME belongs to the class of feature-based linear surrogate models that explain the AI's behavior for individual observations. Notably, "local" refers to the possibility to explain how a certain combination of input features shape the associated, individual prediction.

In a nutshell, it works as follows. LIME first creates artificial, perturbed data points in the local proximity around the instance for which it produces explanations. For every artificial data point, the original "black box" model produces a prediction. Subsequently, LIME fits a linear, intrinsically interpretable model (here: Ridge regression) on the created data set, whereby it weighs artificial data points according to their distance to the real data point. Estimated local coefficients for the input features of the real data point then depict how this very attribute contributes to the overall prediction of the "black box" model. For instance, for a specific male borrower who is highly competitive, LIME might estimate that for this very person being male decreases the likelihood of repayment by 10 %, while his high competitiveness increases the likelihood of repayment by 5 %.

We opted for LIME because its explanations are highly intuitive and straightforward to explain and interpret for lay users. Following the standard approach suggested by Ribeiro et al. (2016a), we visualize explanations graphically using red and green colored bars, respectively depicting a negative or positive contribution of the corresponding characteristic to the GBRF's prediction. The length of bars indicates the quantitative strength of the contribution. For instance, a long red bar indicates that, for the given borrower, the corresponding characteristic is strong evidence against him paying back an investment. A short green bar indicates that, for the given borrower, the corresponding characteristic is weak evidence in favor of him paying back an investment. To avoid biases associated with subjective interpretations of probabilities, we did not display underlying probability values. Instead, we only depict estimated local coefficients as colored bars. We explain to participants in detail how they have to interpret the bars.

[5] Using the standard ten-fold cross-validation, the model achieves an average performance of about 74% accuracy.

Notably, although we use LIME, it more broadly reflects model-agnostic methods that produce local explanations about how individual input factors contribute to given predictions. Instead of LIME, we could also have used local explanations produced by SHAP (Lundberg and Lee 2017). Hence, our results should be interpreted in the light of potential effects associated with local, model-agnostic explanations that, at least partially, rely on intuitive graphical visualizations.

While we use the training set as the basis of our (explainable) GBRF, the player set serves as the representative out-of-sample population of borrowers against which participants in our experiment play. On the player set, the GBRF achieves a performance of 69.8% accuracy, i.e., correctly predicts borrowers' repayment behavior in more than two-thirds of the cases. To determine the outcomes and payoffs for a given investment decision, we match the online study participants' corresponding investment decision with the conditional decision of the field study participant. Notably, to implement an actual strategic setting, we recontact and pay field study participants according to the outcomes of a randomly drawn subset of investment games. We make online study participants explicitly aware of this feature so that they understand that their decisions affect the material well-being of other people as well as their payoff in this study.

Using the participants from the previous field study as borrowers has two advantages. First, due to this procedure borrowers are drawn from the same population as the training data, ensuring that the Gradient Boosted Forest performs reasonably well. Second, it reduces the complexity of the experiment for online participants so that we mitigate fatigue concerns while at the same time maximizing the number of observations we are mainly interested in.

3.3 Stage I

In stage I, participants play ten rounds of the outlined one-shot investment game against different borrowers. For every participant, we randomly draw ten different borrowers without replacement from the player set. This way, we control for order effects. Before participants make their investment decisions they observe the ten characteristics of the borrower they can invest with in the given round. While we fix the order in which we present the characteristics to a given investor across all investment decisions she makes, we randomized the order across investors. We do so to control for order effects while at the same time reducing the cognitive effort associated with processing information to decide. We do not provide intermediary feedback to prevent the development of expertise, idiosyncratic investment strategies, and path dependencies based on the consequences of investment decisions, because such effects might confound our results.

Stage I serves two purposes. First, despite the absence of feedback, participants can familiarize themselves with the investment task for the subsequent stages and form prior beliefs about the relevance of borrower characteristics and their relation to repayment behavior. Second, elicited investment decisions allow us to identify participants' prior choice patterns and thereby developed beliefs about the relationship between borrowers' characteristics and repayment behavior.

3.4 Stage II

In stage II, participants play one investment game against a random borrower from the player set whom they have not encountered in stage I. In contrast to the previous stage, participants can only observe three out of the ten borrower characteristics, before making their investment decision. Participants have to choose the characteristics they prefer to see. Specifically, we ask them to select three distinct characteristics and mark them as first, second, and third choice. They observe the characteristics marked as the first choice before making their investment decision with a probability of 1. They see their second and third choices with a probability of 0.9 and 0.8, respectively. With the corresponding inverse probabilities of 0.1 and 0.2, they instead observe distinct characteristics of the borrower that we randomly draw from the remaining seven characteristics that the participant does not select. We randomly determine the three characteristics participants actually observe according to the outlined probabilities. To ensure incentive compatibility the investment decision in this round is payoff relevant in any case. Again, participants do not receive feedback on the outcome of the game.

3.5 Stage III

Stage III comprises 20 rounds of the investment game against distinct random borrowers from the player set that participants have not encountered before. There is no feedback on game outcomes between rounds. As in stage I, participants observe all of the borrowers' ten personal characteristics before making their investment decision. Additionally, participants also observe the (explainable) AI system's prediction about whether the borrower repays an initial investment.

To reduce potential initial skepticism towards the AI, we explain to participants in detail how the model operates, how it has been trained, and reveal its performance on a representative test set, i.e., we provide global explanations about the AI. Notably, we explicitly inform participants that the model produces the prediction only using the borrowers' ten personal characteristics they also observe. That is, we emphasize that the model does not have access to any additional

information about the borrower. This way, we make sure that participants understand that the AI has no information advantage due to additionally observed signals. Subjects observe a binary prediction that we formulated as an unambiguous text to avoid misinterpretations.[6]

Our between-subject treatment variation is whether or not participants, in addition to the prediction as such, also receive a human-interpretable explanation about the contribution of borrower characteristics to a specific prediction using LIME (Local Interpretable Model-Agnostic Explanation, Ribeiro et al. 2016a). In our treatment condition, participants observe LIME explanations for each borrower characteristic, informing them whether it is evidence for or against the borrower repaying an investment and how strong it is. To avoid confusion, we explain to participants in detail how they should interpret the explanations. By contrast, baseline participants do not see any additional explanation. At this point it is important to understand that participants in both conditions actually interact with the same AI, producing the same predictions for the same borrower. The only difference is that in the treatment, we also provide post-hoc, model-agnostic explanations. We portray our operationalized interface in appendix B.

We measure baseline (treatment) participants' trust in the (explainable) AI's predictive performance for the first and the second ten rounds of investment decisions. In both cases, participants have to guess the share of accurate predictions for the preceding ten rounds. Subjects receive a payoff of 3 MU for every guess that is off by at most 20 percentage points. Hence, we obtain incentive compatible measures of participants' trust in the machine performance.

3.6 Stage IV

Stage IV mirrors stage II. That is, we measure participants' posterior beliefs about the three most decision-relevant borrower characteristics. Notably, we match participants with a borrower they have not encountered in any previous stage. Again we do not provide feedback.

3.7 Stage V

Finally, in stage V, participants play another ten rounds of the investment game without feedback. Notably, participants play against the same ten individuals that they have encountered in stage I. We randomize the order in which participants play against the borrowers from stage I. Participants again only observe borrowers' ten personal characteristics before making their transfer decision, but no AI prediction at all. Letting participants play against the same individuals as in stage I allows us to observe any individual-level changes across the experiment.

[6] If the produced probability that the borrower reciprocates a transfer is greater than 50%, we inform participants that the borrower will most likely repay an initial investment.

After participants have made all ten decisions, the experiment ends with a questionnaire containing items on participants' socio-demographics and social preferences. Participants' answers serve as controls for some of our regression analyses. At the end of the experiment, we inform participants about the outcomes of payoff relevant investment games and their payoffs.

3.8 Experimental summary

Overall, 607 individuals participated in our study (301 Treatment condition and 306 Baseline condition).[7] We run the experiment as an online experiment on the popular and widely used platform *Prolific*. The experiment is implemented using oTree, Python, and HTML. Participants' earnings equal the sum of MU they earn in each stage. In each stage, we match participants' investor decisions with corresponding borrower decisions to determine payoffs according to the previously outlined structure. For stages I, III, and V where participants make multiple investment decisions, we randomly select one of the rounds. Notably, to mitigate concerns about participants not paying attention to displayed information and rush through the investment decisions, they were allowed to submit investment decisions after at least 5 seconds. On average, participants earned \$5.52 (\$4 participation fee; \$1.52 due to actual decisions) and took about 27 minutes to finish the experiment.[8]

4 Results

In accordance with the three main questions we have in mind, we present our results in three parts. First, we analyze how the employment of machine explanations affects participants' weighting of available information by comparing investment decisions between stages I and III. Second, we examine to what extent machine explanation affect participants' preexisting beliefs about the relationship between borrowers' characteristics and repayment behavior. We do so by analyzing changes in participants' decisions across stages I and V. Finally, we study the importance of trust in the explainable AI's predictive performance for belief adjustment processes.

[7] This study was approved by the IRB of the ... (blinded for review) and preregistered at the American Economic Association's registry for randomized controlled trials (AEA RCT Registry. December 07. https://doi.org/10.1257/...) (blinded for review).

[8] For every transfer decision that is ultimately payoff relevant for participants in the experiment, we randomly draw a number between 0 and 20. If the drawn number is equal to 20, we contact and pay the corresponding borrower according to the game's outcome.

4.1 XAI and decision making

We analyze changes in the influences of available information entailed by the provision of machine explanations using regression analyses. Table 1 shows standardized estimates of these analyses. In each regression model, investment decisions serve as the dependent variable. Independent variables are all pieces of information that participants observe and borrowers' unobserved type, which allows us to capture additional borrower types fixed effects. To estimate differences across stages within baseline and treatment conditions (see columns (3) and (6)), we additionally include a stage dummy and corresponding interaction terms. We also conduct a Difference-in-Difference (DiD) analysis (see column (7)) to isolate changes in coefficients purely associated with the provision of machine explanations. In all regressions we include participant and stage fixed effects cluster robust standard errors on the individual $\times$ stage level.[9]

Regression results show that the provision of machine explanations, in addition to predictions, changes the influence of several observed information on investment decisions. First, we find that participants are significantly less likely to follow an explained repayment prediction. In comparison to opaque repayment predictions, participants are 5.2 percentage points (standardized units) less likely adhere to a repayment prediction and invest. Second, while the mere provision of an opaque prediction appears to decrease (most) borrower characteristics' influence on investment decisions, the introduction of machine explanations seems to counteract or reinforce these weight changes, conditional on the characteristic. Specifically, explanations significantly reinforce the influence of the competitiveness, patience, gender, and older siblings by 7.9 ($p < 0.01$), 3.9 ($p < 0.05$), 4.9 ($p < 0.01$), and 2.5 percentage points ($p < 0.1$) per one standard deviation, respectively. By contrast, the influence of agreeableness significantly decreases by 4.7 percentage points per one standard deviation. In the context of the investment decision these results indicate that with explanations, participants are less likely to blindly follow a prediction depicting a borrower to repay an investment. Instead they seem to look at some characteristics more closely before deciding. Specifically, due to explanations, they are less likely to invest in competitive, impatient men (without older siblings) even if they appear warm and considerate.

As one might expect, the direction and magnitude of weight changes largely mirror observed machine explanations. Figure 3 shows the distribution of LIME values for all ten borrower characteristics and their values that treatment participants observed. High positive and negative contributions to the prediction indicate that the associated characteristic value is evidence for or

[9]Thereby we account for the possibility that participants' decision making patterns completely change once we introduce AI decision support.

against repayment, respectively. Machine explanations portray borrowers' repayment likelihood to depend mainly on their competitiveness (strong negative relation), patience (strong positive relation), and gender (strong negative effect for males). Our DiD regression results depict that the introduction of explanations significantly elevates participants weighting of these three characteristics. We detect the strongest increase in the estimate for competitiveness, i.e., the characteristic machine explanations mark as most influential. Changes for patience and gender, which the machine depicts are equally important, are of similar magnitude. For agreeableness, that machine explanations mark as relatively unimportant, we find treatment participants to consider it significantly less intensely.[10]

In sum, our findings so far show that the provision of explanations about how input features contribute to the AI's prediction lead participants to significantly change their weighting of available information in directions suggested by machine explanations. Participants interacting with the explainable AI weigh borrower characteristics marked as most important significantly more, while becoming significantly less likely to follow a "repayment" prediction. In fact, we find that participants are more likely to rubberstamp a prediction when explanations about how the observed borrower characteristics relate to the prediction accompany it. Baseline and treatment participants override the prediction in 22.6 % and 27.8 % of the cases, respectively. The difference is economically (+23 %) and statistically significant ($p < 0.001$, χ^2-test), indicating that explanations cause participants to rely less often on the overall prediction. Notably, beliefs about the prediction accuracy does not significantly differ between opaque and explained predictions (71.8 % and 70.6 % respectively, $p = 0.751$, Wilcoxon rank-sum test), so that changes in the weighting of predictions do not seem to result from lower trust. It thus appears that the provision of machine explanations causes participants to shift their attention away from predictions alone towards additional information.

Result 1: *Machine explanations steer users' attention away from the isolated prediction towards information marked as decision-relevant. This occurs without reducing users' trust in the prediction accuracy.*

These observations suggest that the provision of machine explanations may reduce the re-

[10]Even though machine explanations suggest that having older siblings is the least relevant characteristic to predict repayment behavior, we observe a marginally significant increase in the influence on investment decisions due to explanations. A plausible rationale for this observation is that machine explanations lead participants to consider every borrower characteristic at least to some minimum degree. Hence, machine explanations could significantly increase participants' consideration of a characteristic when predictions steer their attention (strongly) away from it. Providing support for the notion of a minimum attention level, all estimates in column (5) are at least marginally significant, except for agreeableness which is slightly insignificant ($p = 0.2$).

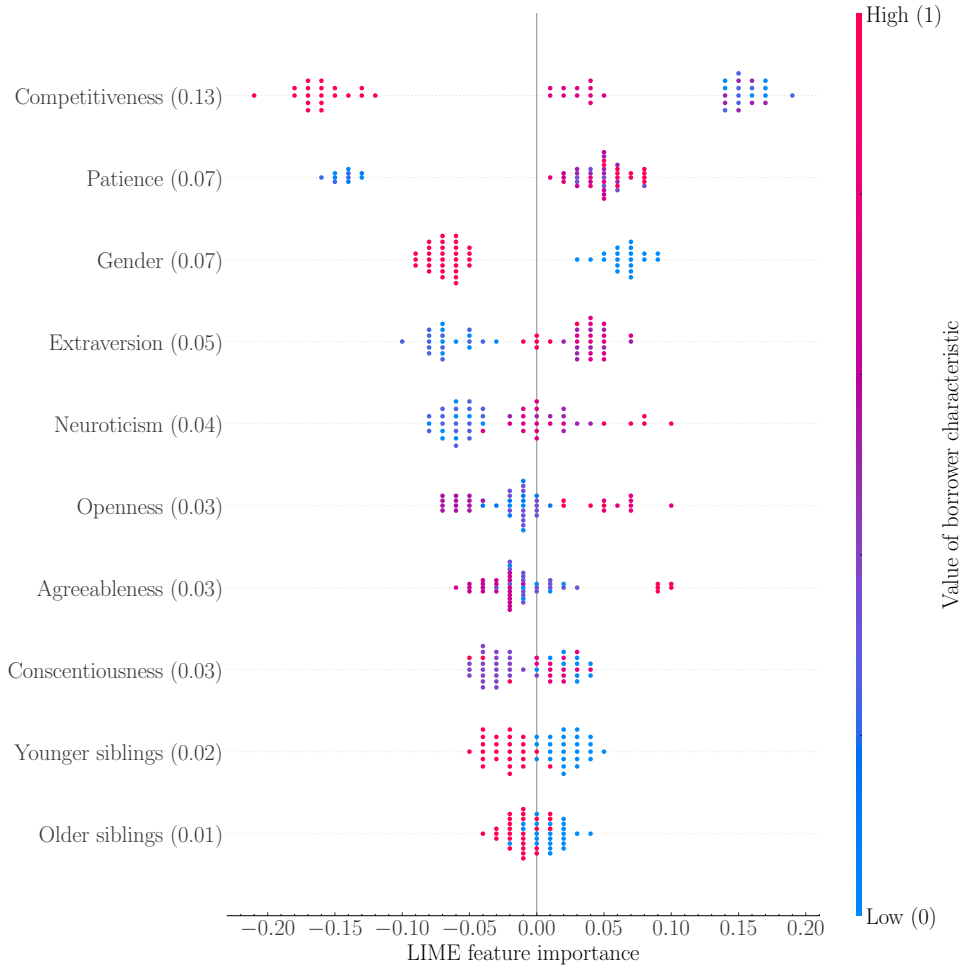

Figure 3: Distribution of LIME values

Notes: Distribution of LIME values, i.e., prediction contribution, for different borrower characteristics. The color of dots indicates the level of the corresponding borrower characteristic. From top to bottom characteristics are in a descending order according to their mean absolute prediction contribution (noted in parentheses).

liance on predictions alone and cause users to process other available information marked as decision-relevant more vigilantly. Hence, as one might hope, machine explanations appear to be an effective means to reduce automation biases (Skitka et al. 2000, 1999), i.e., the in appropriate overreliance on intransparent outputs of AI systems that leads users to ignore other relevant information and impairs decision making. Notably, explanations can induce such a shift in attention without harming participants' overall trust in the predictive performance of the AI.

Our first result demonstrates that the employment of explainability measures have important consequences on how users engage with the AI and weight other available information. However, it remains open whether these effects only occur because users effectively delegate their (real)

decision authority (see, e.g. Aghion and Tirole 1997, Baker et al. 1999) to the explainable AI or if explanations cause belief adjustments about the relation between borrower characteristics and repayment behavior (see, e.g. Hogarth and Einhorn 1992, Epstein et al. 2010, Rabin 2013, Henckel et al. 2021). When the latter is the case, the employment of explainable AI may have much more significant consequences on human behavior than merely shifting users' attention. We examine this notion in the subsequent section.

4.2 XAI and belief updating

At the heart of human behavior lie beliefs about, for instance, the consequences of their decision making. People form and update their beliefs in response to receiving new information, in some way or another (see, e.g. Hogarth and Einhorn 1992, Epstein et al. 2010, Rabin 2013, Henckel et al. 2021). In our study, participants initially form their beliefs about the relationship between borrower characteristics and repayment behavior to make investment decisions in stage I. During stage III, treatment participants observe machine explanations about how borrower characteristics relate to repayment predictions. These explanations effectively constitute new information that may lead treatment participants to adjust their beliefs initially formed in stage I, and thus change their behavior.[11] To detect any ongoing belief updating, we compare participants' investment choices in stages I and V, i.e., before and after they interacted with the opaque or explainable AI. Recall that participants had to make decisions for the same ten borrowers in the two stages, even though in a randomized order. Therefore, we can directly compare their choices for a given borrower across stages. As before, we use (Difference-in-Difference) regression analyses to determine changes in the influence of observed borrower characteristics on investment decisions over the course of the experiment.

Table 2 shows results of our regression analyses. In each regression, the investment decision serves as the dependent variable. In columns (1) and (2), observed borrower characteristics and their unobserved type, a posterior dummy, and corresponding interaction terms as independent variables. Column (3) depicts results of a corresponding Difference-in-Difference regression. We control for participant and stage fixed effects and cluster robust standard errors on the individual $\times$ stage level, to account for the fact that participants' way of decision making before and after interacting with the AI may fundamentally differ. We report standardized coefficients of interaction terms, depicting changes in the weighting of characteristics from prior to posterior decision making.

[11] As we do not provide feedback about the game outcomes, participants are by design unable to learn the true quality of the signals or engage in some version of reinforcement learning. This allows us to draw clear conclusions about the pure effects of machine explanations on beliefs. Additionally, it facilitates the interpretation of our results.

Regression results indicate that the provision of machine explanations leads to selective belief adjustments about the importance of borrower characteristics for investment decisions. On the one hand, we find that the influence of competitiveness, patience, and gender on investment decisions increases after participants interacted with the explainable AI (respectively by 5.2, 2.1, and 2.7 percentage points per one standard deviation). That is, treatment participants do not only put more weight on these three characteristics while they observe machine explanations that mark them as relevant. Instead, they seem to continue to weigh them more, even when they do not have access to the explainable AI anymore. DiD estimates show that the changes associated with observed explanations are statistically significant suggesting that the provision of machine explanations have led participants to reinforce their preexisting beliefs about the relevance of these traits for their investment decision. On the other hand, it appears as if participants are reluctant to adjust strong preexisting beliefs about the relevance of characteristics for which observed machine explanations suggest that their initial weighting is incorrect. Independent of the condition, we do not find the weight of agreeableness to change after participants interacted with the AI. DiD estimates further depict that changes associated with explanations are insignificant. Agreeableness is (one of) the most influential borrower characteristic for both prior and posterior decisions. That is the case even though participants only put low weight, if any, on agreeableness while interacting with the explainable AI. It seems that participants are willing to adhere to machine explanations that contradict strong preconceptions but unwilling to adjust beliefs accordingly so that they would continue to put less weight on it, when they do not have access to the explainable AI anymore.[12]

In sum, findings regarding the adjustment of preexisting beliefs suggest that participants engage in an asymmetric updating. On the one hand, they appear to reinforce prior beliefs when machine explanations provide support for their preconception. On the other hand, we observe that the influence of the initially most important characteristic does not reside after participants observed machine explanations marking it as irrelevant. Notably, an analysis of participants' decisions in stages II and IV, where they had to select the three characteristics they perceive to be most decision-relevant, corroborate these observations (see Appendix A).

Result 2: *Machine explanations cause users to adjust their beliefs about the relevance of bor-*

[12]Note: After participants interacted with the explainable AI that depicts openness as relatively irrelevant (similarly irrelevant as agreeableness), openness becomes significantly less influential for posterior investment decisions. On the first sight, this pattern may suggest that participants are, at least to some degree, willing adjust their beliefs in the light of contradicting information, after all. However, Table 1 shows that their prior weighting of openness is considerably smaller compared to agreeableness for which we observe a reluctance to adjust beliefs ($p < 0.000$, F-test). Therefore, it seems more plausible to assume that the willingness to actually adjust beliefs decreases with the strength of the preexisting beliefs.

rower characteristics in an asymmetric way. Users reinforce preexisting beliefs that machine explanations substantiate but do not abandon strong prior beliefs that machine explanations depict as incorrect.

How can we account for this asymmetry in adjusting beliefs based on machine explanations? One plausible rationale is that the asymmetry is the manifestation of the confirmation bias, a predilection to search for, favor, process, and even remember information in a way that is favorable to already held beliefs (see e.g., Edwards and Smith 1996, Nickerson 1998, Ditto and Lopez 1992, Park et al. 2013). Specifically, the finding that participants' weighting of agreeableness reverts to very high prior levels while the influence of competitiveness, patience, and gender remains elevated, once they do not observe explanations anymore, might reflect the biased recall of information conditional on whether they support or contradict prior beliefs (see e.g., Frost et al. 2015). It is worth noting that the confirmation bias has been found to play a major role for interacting with human advisors (Agnew et al. 2018b,a). In our case, the confirmation bias is able to reconcile the observations that participants adhere to explanations with regard to agreeableness as long as they interact with the explainable AI, but seem to *cherry pick* explanations when it comes to adjusting their beliefs, i.e., learn from explanations. Put differently, users seem to harness explanations not so much to expand their knowledge, but to support their preexisting perceptions. Providing some support for this assertion, additional regression analyses (see 6 in appendix) depict that only for participants who do not initially rank agreeableness as most important in stage II the coefficient capturing weight changes is negative. The coefficient for individuals who rank agreeableness first is, if anything, positive. Notably, while the difference between the coefficients is in support of our argumentation, it is statistically insignificant.

As the final step of our analyses, we take a closer look at participants' readiness to adjust their beliefs according to observed machine explanations. Specifically, we examine whether there exist treatment heterogeneities conditional on participants' trust in the predictive performance of the AI. Put differently, we ask whether high trust in the AI's prediction accuracy encourages participants to adjust their beliefs.

4.3 Belief updating and trust

Previous studies have shown that trust in the predictive performance of AI decision aid fosters the adoption and use of the technology (Komiak and Benbasat 2006, Logg et al. 2019, Shin et al. 2020, Glikson and Woolley 2020). Additionally, evidence from the advice taking literature

suggests that people's propensity to discount new information in favor of prior beliefs decreases with the perceived expertise of the person (or system) giving advice (Yaniv 2004, Pilditch et al. 2020). From this perspective, it seems plausible to suspect that trust also plays a moderating role when it comes to learning from machine explanations. If participants trust the AI to make accurate predictions, they may be more likely to adjust prior beliefs according to the machine explanations observed.

We harness participants' incentivized guesses about the prediction accuracy in stage III. That is, we consider the cognitive, rather technical aspect of trust in the "expertise" of the AI system (Komiak and Benbasat 2006). To determine the role of trust in the AI's predictive performance for belief adjustment, we perform a median split based on participants' guess about the prediction accuracy. We refer to participants who estimate that the prediction accuracy is at least 75 % as trusting, and the other half as non-trusting types.[13] We examine changes in the weight participants put on observed borrower characteristics across stages I and V. As before, significant changes reflect that participants weigh observed information differently after interacting with the AI. We conduct separate regression analyses for the two different types of participants. Table 3 depicts estimates of the weight changes across stages I and V. In each model, participants' investment decisions serve as dependent variables, while observed borrower characteristics, unobserved borrower types, a posterior dummy, and corresponding interaction terms serve as independent variables. We include participant and stage fixed effects and cluster robust standard errors on the individual × stage level. Columns (1) and (2) portray weight changes for non-trusting baseline and treatment participants, respectively. We report corresponding Difference-in-Difference estimates in column (3). Columns (4), (5), and (6) show results for trusting types.

Comparing the changes in the influence of observed borrower characteristics on participants' investment decisions for different types indicates that only trusting types are willing to learn (asymmetrically), from machine explanations. Non-trusting types who observe machine explanations only adjust their beliefs significantly with regard to openness (weight on openness decreases). Otherwise, they do not seem to engage in belief updating (see column (2)). By contrast, trusting treatment participants exhibit adjustments in the weight they put on borrower characteristics (see column (5)). For the most part, these changes seem to correspond to observed machine explanations. Notably, trusting types exhibit the same kind of confirmation bias we detected on the aggregate level. That is, even though these participants have high confidence in the explainable AI's prediction accuracy, they appear unwilling to learn from it,

[13]Note: 75 % is the median for both the baseline and treatment condition.

not to put weight on the observed agreeableness.

In sum, we find trust in the predictive performance of the AI to play an important role when it comes to learning from machine explanations. Hence, it does not merely play a role when it comes to inciting participants to follow and adhere to AI outputs (see e.g., Glikson and Woolley 2020), but also for explainable AI's capacity to influence the way users understand their (decision-)environment. Notably, analyses of different types' investment decisions while interacting with the AI reveal that non-trusting treatment participants do not only rely less on the overall prediction compared to trusting ones, but also put significant weight on agreeableness. Otherwise, however, their weighting of information is very similar to trusting participants. That is, while trust in the prediction performance of the explainable AI does not seem to spur the adjustment of strong prior beliefs, it appears to influence participants' adherence to explanations that contradict these priors—at least while they observe machine explanations.

Result 3: *Trust in the predictive performance of the explainable AI is an important prerequisite for users to adjust beliefs in the direction of observed machine explanations. However, even users with high confidence in the prediction accuracy do not adapt strong priors that machine explanations oppose.*

The observation that trust in the predictive performance of the XAI is in line with previous findings that people's likelihood to listen to advice by another human increases with the advisor's perceived task-relevant expertise (Sniezek et al. 2004, Bonaccio and Dalal 2006). This observation suggests that the identified relation between people's propensity to rely on (and internalize) advice from another person apply, at least partially, also in human-XAI collaborations.

5 Discussion and implications

In a nutshell, we report three main findings revealing that feature-based XAI causally affects important cognitive processes of human users, i.e., information processing and belief updating. First, we demonstrate that the provision of machine explanations causally changes participants' cognitive processing of available information. Second, we find that the provision of machine explanations entails an asymmetric adjustment of human beliefs that is favorable to strong predilections. Third, (cognitive) trust in the prediction accuracy of the AI is a vital prerequisite for learning from machine explanations, however, it does not prevent asymmetric belief adjustments.

Reported results on the asymmetric belief adjustment have implications for policymakers advocating higher explainability and transparency in AI applications. According to our findings, the provision of feature-based explanations reinforces preexisting beliefs that are in line with the correlations between features and target variables detected by the AI system in the data. This emphasizes the importance not to include sensitive information such as gender or race into AI systems if there are strong correlations with the label and the system eventually has to provide (feature-based) explanations for predictions. Consider, for instance, predictive policing, i.e., predictive AI systems that help allocate police across a city to best prevent crime (Ferguson 2017). According to our results, the provision of explanations to officers that the share of specific ethnicities in certain neighborhoods is strong evidence for a crime occurring may have foster preexisting racial biases. Indeed, in many domains in which AI plays a major role, racial or gender biases persist (e. g., in online hiring (Chan and Wang 2018) or crowdfunding (Burtch and Chan 2019)). Hence, regulatory efforts that aim to make AI systems' predictions more transparent should go hand in hand with measures ensuring that these systems do not include specific sensitive information (correlating with the target measure).

From a different point of view, our results imply that people with a high level of trust in the predictive performance of AI systems are willing to learn from machine explanations and change their behavior. This finding is important from the perspective that a largely untapped potential to enhance economic efficiency is the transfer of domain knowledge from AI systems to human users (see e.g., Teso and Hinz 2020). Our results imply that feature-based machine explanations and trust are an important prerequisite for harnessing the broader potential of AI systems teaching humans, i.e., "machine teaching" (Abdel-Karim et al. 2020). Against this background, organizations that intend to transfer domain knowledge extracted by AI system from (Big) Data to employees, may be well advised to (i) employ feature-based explainability techniques and (ii) foster users' trust in the system's predictive performance. In order to achieve the latter, future research should focus on how this trust is built, similar to recent work exploring how trust is generated to improve economic efficiency in human-human collaboration (Susarla et al. 2020). As a final word of caution on this implication, we want to point out that there may also be a dark side to learning from explanations. There is the danger that ill-meaning third parties harness explainable AI to manipulate people's fundamental belief structures. In a worst case scenario, machine explanations may be misused to spread specific ideological and discriminatory views. From this perspective, our results complement warnings recently asserted in the literature (Lakkaraju and Bastani 2020, Liel and Zalmanson 2020).

6 Conclusion

This paper explores the interplay of feature-based XAI, information processing, and human beliefs. We develop a novel incentivized experimental protocol that allows us to circumvent severe endogeneity concerns we would encounter in a field setting and, thus, identify causal effects. Our paper is the first to provide causal empirical evidence on feature-based XAI's potential to fundamentally change important cognitive processes: information processing and belief adjustment.

In light of the gravity of this potential, there is need for future research in several directions. For instance, it is important to better understand the origins of the asymmetric learning from machine explanations. What are moderating factors? Are there effective ways to alleviate the seeming occurrence of the confirmation bias when it comes to learning from AI? Answering these questions will be a vital step to tap into the potential of explainable AI effectively teaching humans new domain knowledge extracted from Big Data (Abdel-Karim et al. 2020, Bauer et al. 2021).

Another fruitful avenue is to examine whether other XAI techniques, e.g., global explanations or counterfactual explanations, evoke similar cognitive responses so that we have a better idea about the consequences of employing specific XAI systems. Do different types of explanations affect beliefs differently, or not at all? Are users more willing to learn new knowledge from an XAI system that provides counterfactuals? Providing an answer to this question is important to inform organizational decisions about which method to employ.

As a concluding remark, we want to point out that one should interpret our results as evidence that puts assertions about XAI being the silver bullet that solves all of AI systems' (black box) problems into perspective. As with every other technological innovation, XAI comes with merits and problems. Specifically, in our study, the provision of state-of-the-art feature-based explanations affects users' beliefs in an asymmetric way making them potentially vulnerable to manipulation by ill-meaning third parties who can influence machine explanations. Additionally, in our specific setting, the use of feature-based explanations decreases the economic efficiency of decision making. That seems to be the case because XAI users rubberstamp the overall machine prediction in favor of individual, explained features too often. Overall, they are significantly less likely to make the payoff maximizing decisions than opaque AI system users. (respective average shares of payoff maximizing investment decision: 57.5 % and 63.1 %; $p < 0.001$; χ^2-test).[14] While we acknowledge that these results inextricably link to our specific

[14]Note: on average, always following the prediction would have resulted in the payoff maximizing decision in 69.3 % of the cases.

design choices, they nonetheless show that, once again, the human factor can be the weakness of even well-designed XAI systems. This flaw in the system can ultimately impede its efficacy in alleviating the black box problems and possibly creating new, unanticipated problems. Hence, regulations such as Europe's recent proposal for an Artificial Intelligence Act may benefit from not merely stipulating the provision of "meaningful" explanations in whatever form. Additionally demanding a specific presentation of explanations (or their exclusion) may be worthwhile as well.

	Baseline			Treatment			
Dep. variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Making an	Without	With	Δ	Without	With	Δ	Δ
investment	AI	opaque AI	(1)-(2)	AI	explainable AI	(4)-(5)	(3)-(6)
Agreeableness	0.081***	0.049***	-0.032**	0.084***	0.008	-0.077***	-0.041**
	(0.010)	(0.007)	(0.013)	(0.011)	(0.006)	(0.012)	(0.017)
Competitiveness	-0.039***	-0.019***	0.020	-0.034***	-0.097***	-0.062***	-0.079***
	(0.012)	(0.007)	(0.014)	(0.011)	(0.010)	(0.015)	(0.020)
Conscientiousness	0.004	0.017***	0.014	0.018**	0.018***	0.001	-0.014
	(0.009)	(0.006)	(0.011)	(0.008)	(0.006)	(0.010)	(0.015)
Extraversion	0.027***	0.012**	-0.015	0.033***	0.031***	-0.002	0.014
	(0.009)	(0.006)	(0.011)	(0.010)	(0.006)	(0.012)	(0.016)
Patience	0.031***	0.005	-0.027**	0.037***	0.048***	0.011	0.039**
	(0.008)	(0.007)	(0.011)	(0.008)	(0.007)	(0.011)	(0.015)
Openness	0.024***	0.011**	-0.013	0.029***	0.009	-0.020*	-0.004
	(0.009)	(0.005)	(0.010)	(0.009)	(0.006)	(0.010)	(0.015)
Gender (male)	-0.039***	-0.009	0.030**	-0.015	-0.033***	-0.018	-0.049***
	(0.010)	(0.007)	(0.012)	(0.011)	(0.007)	(0.013)	(0.017)
Neuroticism	-0.030**	-0.010	0.019	-0.011	0.013*	0.025*	0.001
	(0.012)	(0.007)	(0.014)	(0.011)	(0.007)	(0.013)	(0.018)
Younger sibl. (yes)	-0.004	0.004	0.009	-0.020**	-0.015***	0.005	-0.006
	(0.009)	(0.005)	(0.011)	(0.008)	(0.005)	(0.010)	(0.014)
Older sibl. (yes)	0.028***	-0.000	-0.028***	0.024***	0.024***	0.000	0.025*
	(0.009)	(0.005)	(0.010)	(0.009)	(0.005)	(0.010)	(0.014)
Repayment prediction		0.224***			0.172***		-0.052***
		(0.012)			(0.011)		(0.017)
N	3060	6120	9180	3010	6020	9030	18210
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Adj. R^2	0.315	0.423	0.389	0.382	0.374	0.383	0.387

Table 1: Regression analyses: Weighing of borrower characteristics

Notes: Regression analyses with participant and stage fixed effects. In each regression, the investment decision serves as the dependent variable. As independent variables, we include observed borrower characteristics, borrowers' unobserved types, and, if applicable, the observed prediction. Reported estimates are standardized. In columns (1), (2), (4), and (5), we cluster robust standard errors on the participant level; in the remaining columns, on the participants × stage level. Column (7) depicts Difference-in-Difference regressions. We denote significance levels as $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Dep. variable:	(1)	(2)	(3)
Making an investment	Opaque AI	Explainable AI	Diff.-in-Diff.
Agreeableness	0.016	-0.005	-0.019
*Posterior	(0.015)	(0.015)	(0.021)
Competitiveness	-0.006	-0.052***	-0.044*
*Posterior	(0.017)	(0.018)	(0.024)
Conscentiousness	0.033***	0.017	-0.017
*Posterior	(0.013)	(0.012)	(0.018)
Extraversion	0.018	-0.010	-0.027
*Posterior	(0.014)	(0.014)	(0.020)
Patience	-0.013	0.021*	0.037**
*Posterior	(0.013)	(0.013)	(0.018)
Openness	0.000	-0.030**	-0.028
*Posterior	(0.013)	(0.013)	(0.018)
Gender	0.019	-0.027*	-0.047**
*Posterior	(0.014)	(0.015)	(0.021)
Neuroticism	0.045***	0.005	-0.044**
*Posterior	(0.016)	(0.016)	(0.022)
Younger sibl.	-0.004	0.002	0.006
*Posterior	(0.014)	(0.012)	(0.018)
Older sibl.	0.005	-0.005	-0.014
*Posterior	(0.013)	(0.013)	(0.018)
N	6120	6020	12140
p	0.000	0.000	0.000
Adj. R^2	0.316	0.355	0.335

Table 2: Regression analyses: Influence of borrower characteristics

Notes: Regressions analyses including participant and stage fixed effects. In each regression, the investment decision serves as the dependent variable. As independent variables, we include observed borrower characteristics, borrowers' unobserved types, a posterior dummy, and their interaction terms. Reported estimates are standardized. We cluster robust standard errors on the participant × stage level to account for the fact that participants' way of decision making before and after interacting with the AI may fundamentally differ. We denote significance levels as $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Dep. variable:	Non-trusting types			Trusting types		
Making an	(1)	(2)	(3)	(4)	(5)	(6)
investment	ΔBaseline	ΔTreatment	Δ (1)-(2)	ΔBaseline	ΔTreatment	Δ (4)-(5)
Agreeableness	0.013	0.029	0.015	0.019	-0.034	-0.048
*Posterior	(0.023)	(0.021)	(0.031)	(0.021)	(0.021)	(0.030)
Competitiveness	-0.002	-0.004	-0.003	-0.008	-0.091***	-0.080**
*Posterior	(0.025)	(0.027)	(0.036)	(0.022)	(0.024)	(0.032)
Conscientiousness	0.040**	0.003	-0.037	0.030*	0.029*	0.001
*Posterior	(0.020)	(0.019)	(0.027)	(0.016)	(0.016)	(0.023)
Extraversion	-0.010	0.005	0.015	0.044**	-0.019	-0.061**
*Posterior	(0.020)	(0.021)	(0.029)	(0.019)	(0.020)	(0.027)
Patience	-0.040**	-0.017	0.023	0.011	0.053***	0.042*
*Posterior	(0.019)	(0.018)	(0.026)	(0.018)	(0.018)	(0.025)
Openness	0.001	-0.032*	-0.037	0.000	-0.025	-0.021
*Posterior	(0.020)	(0.020)	(0.028)	(0.017)	(0.016)	(0.024)
Gender	0.003	-0.022	-0.026	0.028	-0.028	-0.060**
*Posterior	(0.022)	(0.022)	(0.031)	(0.018)	(0.020)	(0.027)
Neuroticism	-0.007	-0.008	0.001	0.090***	0.014	-0.083***
*Posterior	(0.023)	(0.024)	(0.033)	(0.022)	(0.022)	(0.030)
Younger sibl.	0.003	0.010	0.008	-0.009	-0.005	0.001
*Posterior	(0.021)	(0.018)	(0.028)	(0.018)	(0.016)	(0.024)
Older sibl.	0.002	0.014	0.017	0.008	-0.021	-0.035
*Posterior	(0.019)	(0.018)	(0.026)	(0.018)	(0.018)	(0.025)
N	2780	2800	5580	3340	3220	6560
p	0.000	0.000	0.000	0.000	3220.000	0.000
Adj. R^2	0.298	0.352	0.326	0.322	0.367	0.344

Table 3: Regression analyses: Treatment heterogeneities

Notes: Regressions analyses with participant and stage fixed effects. In each regression, the investment decision serves as the dependent variable. As independent variables, we include observed borrower characteristics, unobserved borrower types, a posterior dummy, and corresponding interaction terms. Columns (1)-(3) show results for the subsample of non-trusting types. Columns (4)-(6) show results for trusting types. Reported estimates are standardized. We cluster robust standard errors on the individual × stage level. Columns (3) and (6) depict Difference-in-Difference regression results, respectively showing the estimated difference between columns (1) and (2) or (4) and (5). We denote significance levels as $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

References

Benjamin M Abdel-Karim, Nicolas Pfeuffer, Gernot Rohde, and Oliver Hinz. How and what can humans learn from being in the loop? *German Journal of Artificial Intelligence*, 34(2):199–207, 2020.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6:52138–52160, 2018.

Marc TP Adam, Jan Krämer, and Marius B Müller. Auction fever! How time pressure and social competition affect bidders' arousal and bids in retail auctions. *Journal of Retailing*, 91(3):468–485, 2015.

Pär J Ågerfalk. Artificial intelligence as digital agency. *European Journal of Information Systems*, 29 (1):1–8, 2020.

Philippe Aghion and Jean Tirole. Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29, 1997.

Julie R Agnew, Hazel Bateman, Christine Eckert, Fedor Iskhakov, Jordan Louviere, and Susan Thorp. First impressions matter: An experimental investigation of online financial advice. *Management Science*, 64(1):288–307, 2018a.

Julie R Agnew, Hazel Bateman, Christine Eckert, Fedor Iskhakov, Jordan J Louviere, and Susan Thorp. Learning and confirmation bias: Measuring the impact of first impressions and ambiguous signals. *UNSW Business School Research Paper, Wharton Pension Research Council Working Paper*, (2018-08), 2018b.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

George Baker, Robert Gibbons, and Kevin J Murphy. Informal authority in organizations. *Journal of Law, Economics, and Organization*, 15(1):56–73, 1999.

Kevin Bauer, Oliver Hinz, Wil van der Aalst, and Christof Weinhardt. Expl(AI)n it to me–explainable AI and information systems research. *Business & Information Systems Engineering*, 63(2):79–82, 2021.

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

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

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.

Anol Bhattacherjee and G Premkumar. Understanding changes in belief and attitude toward information technology usage: A theoretical model and longitudinal test. *MIS Quarterly*, pages 229–254, 2004.

Ronit Bodner and Drazen Prelec. Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions*, 1(105):26, 2003.

Silvia Bonaccio and Reeshad S Dalal. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2):127–151, 2006.

Gordon Burtch and Jason Chan. Investigating the relationship between medical crowdfunding and personal bankruptcy in the united states: Evidence of a digital divide. *MIS Quarterly*, pages 237–262, 2019.

Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239, 2020.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *International Conference on Healthcare Informatics*, 2015.

Colin F Camerer and Robin M Hogarth. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1):7–42, 1999.

Noah Castelo, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019.

Jason Chan and Jing Wang. Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science*, 64(7):2973–2994, 2018.

Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243 (4899):1668–1674, 1989.

Jan De Spiegeleer, Dilip B Madan, Sofie Reyners, and Wim Schoutens. Machine learning for quantitative finance: Fast derivative pricing, hedging and fitting. *Quantitative Finance*, 18(10):1635–1643, 2018.

Jasbir S Dhaliwal and Izak Benbasat. The use and effects of knowledge-based system explanations: Theoretical foundations and a framework for empirical evaluation. *Information Systems Research*, 7(3):342–362, 1996.

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

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3): 1155–1170, 2018.

Peter H Ditto and David F Lopez. Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63(4):568, 1992.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. In *arXiv:1702.08608*, 2017.

Kari Edwards and Edward E Smith. A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, 71(1):5, 1996.

Larry G Epstein, Jawwad Noor, Alvaro Sandroni, et al. Non-bayesian learning. *The BE Journal of Theoretical Economics*, 10(1):1–20, 2010.

Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 43–52, 2020.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

EU. Proposal for a regulation EU of the european parliament and of the council of 21 April 2021, laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. *Official Journal of the European Union*, L 119, 2021.

Ernst Fehr and Urs Fischbacher. The nature of human altruism. *Nature*, 425(6960):785–791, 2003.

Shi Feng and Jordan Boyd-Graber. What can AI do for me? Evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 229–239, 2019.

Andrew Guthrie Ferguson. Policing predictive policing. *Wash. UL Rev.*, 94:1109, 2017.

Peter Frost, Bridgette Casey, Kaydee Griffin, Luis Raymundo, Christopher Farrell, and Ryan Carrigan. The influence of confirmation bias on memory and source monitoring. *The Journal of general psychology*, 142(4):238–252, 2015.

GDPR. Regulation EU 2016/679 of the european parliament and of the council of 27 april 2016, article 22. *Official Journal of the European Union L 119*, 59, 2016.

Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.

Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, pages 497–530, 1999.

William M Grove and Martin Lloyd. Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology*, 115(2):192, 2006.

William M Grove and Paul E Meehl. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, 2(2):293, 1996.

Junius Gunaratne, Lior Zalmanson, and Oded Nov. The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, 35(4):1092–1120, 2018.

Timo Henckel, Gordon D Menzies, Peter G Moffatt, and Daniel J Zizzo. Belief adjustment: A double hurdle model and experimental evidence. *Experimental Economics*, pages 1–42, 2021.

Mitchell Hoffman, Lisa B Kahn, and Danielle Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

Robin M Hogarth and Hillel J Einhorn. Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, 24(1):1–55, 1992.

Izak Benbasat Ji-Ye Mao. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems*, 17(2):153–179, 2000.

Zhenhui Jiang and Izak Benbasat. Virtual product experience: Effects of visual and functional control of products on perceived diagnosticity and flow in electronic shopping. *Journal of Management Information Systems*, 21(3):111–147, 2004.

Zhenhui Jiang and Izak Benbasat. The effects of presentation formats and task complexity on online consumers' product understanding. *MIS Quarterly*, pages 475–500, 2007.

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

Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *European Conference on Information Systems (ECIS)*, 2020.

Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, forthcoming, 2021.

Daniel Kahneman and Amos Tversky. On the psychology of prediction. *Psychological review*, 80(4):237, 1973.

René F Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395, 2016.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

Sherrie YX Komiak and Izak Benbasat. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, pages 941–960, 2006.

Arie W Kruglanski. Motivated social cognition: Principles of the interface. *Social Psychology: Handbook of Basic Principles*, pages 493–520, 1996.

Himabindu Lakkaraju and Osbert Bastani. "How do i fool you?" Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, 2019.

Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

Yotam Liel and Lior Zalmanson. What if an AI told you that 2+ 2 is 5? Conformity to algorithmic recommendations. In *International Conference on Information Systems (ICIS)*, 2020.

Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *arXiv:1705.07874*, 2017.

Paul E Meehl. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press, 1954.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. In *arXiv:1802.00682*, 2018.

Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. In *arXiv:1909.09223*, 2019.

Dilek Önkal, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22(4):390–409, 2009.

Andrés Páez. The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, 29 (3):441–459, 2019.

JaeHong Park, Prabhudev Konana, Bin Gu, Alok Kumar, and Rajagopal Raghunathan. Information valuation and confirmation bias in virtual communities: Evidence from stock message boards. *Information Systems Research*, 24(4):1050–1067, 2013.

Toby D Pilditch, Jens K Madsen, and Ruud Custers. False prophets and cassandra's curse: The role of credibility in belief updating. *Acta Psychologica*, 202:102956, 2020.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52, 2021.

Andrew Prahl and Lyn Van Swol. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702, 2017.

Matthew Rabin. Incorporating limited rationality into economics. *Journal of Economic Literature*, 51 (2):528–43, 2013.

Matthew Rabin and Joel L Schrag. First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, 114(1):37–82, 1999.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016a.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. In *arXiv:1606.05386*, 2016b.

Drew Roselli, Jeanna Matthews, and Nisha Talagala. Managing bias in AI. In *Companion Proceedings of The World Wide Web Conference*, pages 539–544, 2019.

Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, 2019.

Donghee Shin, Bouziane Zaid, and Mohammed Ibahrine. Algorithm appreciation: Algorithmic performance, developmental processes, and user interactions. In *International Conference on Communications, Computing, Cybersecurity, and Informatics*, pages 1–5, 2020.

Linda J Skitka, Kathleen L Mosier, and Mark Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999.

Linda J Skitka, Kathleen Mosier, and Mark D Burdick. Accountability and automation bias. *International Journal of Human-Computer Studies*, 52(4):701–717, 2000.

Paul Slovic and Sarah Lichtenstein. Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6(6): 649–744, 1971.

Janet A Sniezek, Gunnar E Schrah, and Reeshad S Dalal. Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, 17(3):173–190, 2004.

Anjana Susarla, Martin Holzhacker, and Ranjani Krishnan. Calculative trust and interfirm contracts. *Management Science*, 66(11):5465–5484, 2020.

Stefano Teso and Oliver Hinz. Challenges in interactive machine learning. *German Journal of Artificial Intelligence*, 34(2):127–130, 2020.

Viswanath Venkatesh and Fred D Davis. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2):186–204, 2000.

Viswanath Venkatesh, James YL Thong, and Xin Xu. Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, pages 157–178, 2012.

Giulia Vilone and Luca Longo. Explainable artificial intelligence: A systematic review. In *arXiv:2006.00093*, 2020.

Weiquan Wang and Izak Benbasat. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23(4):217–246, 2007.

Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. How do visual explanations foster end

users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 189–201, 2020.

Ilan Yaniv. Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1):1–13, 2004.

Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP Conference on Human-Computer Interaction*, pages 23–39, 2017.

Florian Zimmermann. The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–61, 2020.

A Supplementary Material

A.1 Prior field study

We collected this data in an incentivized field study that we conducted at a large German university over three years (2016–2019). 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 Euro. Participants could either keep the 10 Euro for themselves or transfer them to their opponent. Whenever one player transferred her 10 Euro, we doubled the amount so that the other player received 20 Euro. 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.

	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.	Patience	(0,1)
8.	Gender	Male=1, Female=0
9.	Person has younger siblings	Yes=1, No=0
10.	Person has older siblings	Yes=1, No=0

Table 4: Features used to train the Machine Learning Model.

Notes: We normalized the scale of numeric items for training and prediction processes.

A.2 Additional results on belief updating

We start comparing participants' selection of the three most decision-relevant characteristics before and after they interacted with the AI. For each borrower characteristic, Figure 4 shows the share of investors who select it among the three most relevant characteristics for their investment decision.[15] Different colored bars represent participant shares before (prior) and after (posterior) participants engaged with the AI. Panel (i) and (ii), respectively, portray baseline and treatment results.

[15]Note: For ease of interpretation we aggregate the ordinal ranking decision so that we consider whether a characteristic has been included in the selection or not.

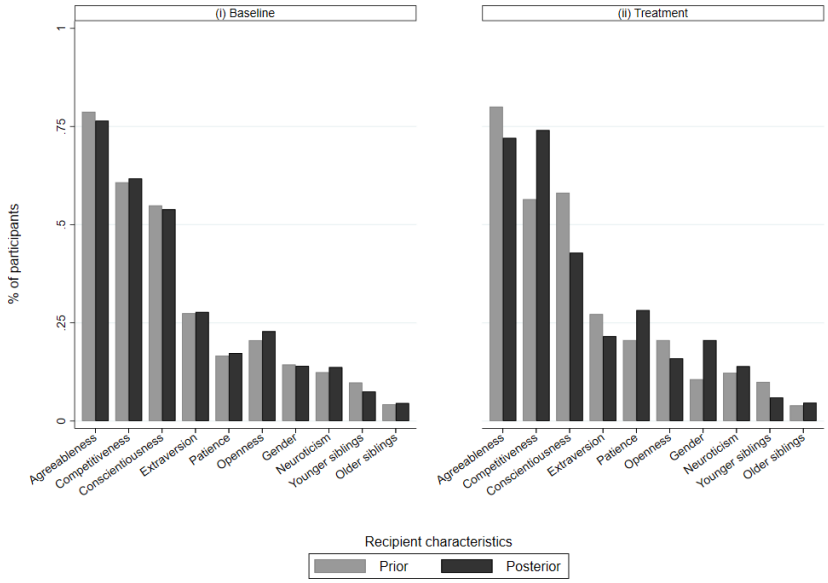

Figure 4: Share of participants ranking characteristics among three most relevant

Notes: Prior and posterior shares of participants who ranked a given characteristic among three most relevant for making the investment decision. Different panels show results for baseline and treatment participants.

Figure 4 indicates that the provision of machine explanations not only impacts immediate decisions. Instead, explanations may alter aggregate beliefs about the decision-relevance of borrower characteristics. Prior and posterior beliefs for participants observing opaque predictions are almost identical (see panel (i)). Before and after interacting with the AI, baseline participants belief agreeableness (prior: 78.8 %, posterior: 76.5 %), competitiveness (prior: 60.8 %, posterior: 61.7 %), and conscientiousness (prior: 54.9 %, posterior: 53.9 %) to be the most informative characteristics, by far. While treatment participants also believe these three characteristics to be most informative at both points in time, panel (ii) depicts some notable changes in the aggregate distribution. On the one hand, the share of participants who believe competitiveness, patience, or gender to be most informative increases from 56.5 % to 74.2 %, 20.6 % to 28.2 %, and 10.6 % to 20.6 %, respectively. On the other hand, the shares of participants selecting conscientiousness decreases from 58.1 % to 42.9 %.

Difference-in-difference (DiD) regression analyses show that these changes are statistically significant (see Table 5). DiD estimates suggest that observing machine explanations leads treatment participants to significantly adjust their beliefs about the decision-relevance of borrowers' competitiveness, conscientiousness, patience, openness, and gender (see columns (2),

Dep. variable: Selecting charachteristic	(1) Agreeab leness	(2) Competit iveness	(3) Conscien tiousness	(4) Extra version	(5) Pati ence	(6) Open ness	(7) Gender	(8) Neuro ticism	(9) Younger. siblings	(10) Older siblings
Treatment	0.010	-0.038	0.011	-0.000	0.045	0.014	-0.036	0.003	-0.004	-0.005
	(0.032)	(0.040)	(0.041)	(0.036)	(0.032)	(0.033)	(0.028)	(0.027)	(0.025)	(0.016)
Posterior	-0.023	0.010	-0.010	0.003	0.007	0.023	-0.003	0.013	-0.023	0.003
	(0.026)	(0.026)	(0.027)	(0.027)	(0.024)	(0.023)	(0.018)	(0.022)	(0.016)	(0.013)
Treatment *Posterior	-0.057	0.166***	-0.143***	-0.060	0.070*	-0.069**	0.103***	0.004	-0.017	0.003
	(0.037)	(0.039)	(0.040)	(0.040)	(0.038)	(0.032)	(0.029)	(0.031)	(0.024)	(0.019)
Constant	0.657***	0.753***	0.649***	0.216*	0.123	0.195*	0.114	0.020	0.126	0.147**
	(0.109)	(0.129)	(0.132)	(0.115)	(0.097)	(0.106)	(0.102)	(0.081)	(0.085)	(0.069)
N	1214	1214	1214	1214	1214	1214	1214	1214	1214	1214
p	0.000	0.000	0.000	0.069	0.034	0.031	0.110	0.015	0.071	0.133
R^2	0.076	0.044	0.044	0.029	0.032	0.036	0.019	0.039	0.025	0.054

Table 5: Regression analyses: Selection of characteristics

Notes: Random effects GLS regression analyses. In each regression, selecting the corresponding characteristic among top 3 in stages II and IV serves as the dependent variable. As independent variables, we include a treatment dummy variable, a dummy indicating the posterior decisions, and an interaction term. We additionally include control variables for participants' social preferences and socio-demographics. We cluster robust standard errors on the individual level to account for the panel data structure, i.e., distinct participants who repeatedly made their investment decisions for different borrowers. We denote significance levels as $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

(3), (5), (6), and (7)).

A comparison of aggregate belief adjustments for the three initially most pronounced beliefs with machine explanations treatment participants observed (see Figure 3) provides some evidence that belief updating depends on the strength of prior beliefs and their alignment with explanations. For agreeableness and conscientiousness, machine explanations and initial human priors are at odds. While the machine depicts these characteristics similarly irrelevant to investment decisions, participants on average believe it to be among the most decision-relevant information. Here we only observe significant belief adjustments for conscientiousness, which participants on average initially rank significantly less often as the most important characteristic (43.2 % vs. 17.9 % of the cases, $p < 0.001$, χ^2-test). Put differently, explanations do not incite participants to abandon the by far most pronounced initial prior. Concerning competitiveness, for which machine explanations and strong initial priors are aligned, we observe a significant reinforcement of the perceived decision-relevance so that it becomes the most decision-relevant information for the majority of treatment participants, after interaction with the machine (48.5 %).[16]

[16]Concerning gender, patience, and openness, which the majority believes not to be among the most relevant characteristics, i.e., rather weak priors on the aggregate level, we observe significant belief adjustments in the direction of the explanation. Notably, for gender, prior and posterior beliefs may be subject to additional self-signaling concerns. This refers to decisions that, at least partially, aim at establishing a desirable view of oneself (Bodner and Prelec 2003, Bénabou and Tirole 2011). In line with this notion, participants may initially consider

These aggregate level results are indicative that explainable AI can have permanent effects on human behavior, while this is not true for an opaque AI. It seems that it is explanations that can lead participants to adjust their beliefs about the individual decision-relevance of borrower characteristics.

basing an investment decision on gender to be a 'taboo'. However, once the AI provides explanations that gender is a relevant factor, they might feel allowed to include this information in their decision. As our experimental design cannot provide further clarification on this issue, an investigation of such an effect presents a fruitful avenue for future research.

A.3 Confirmation bias

Dep. variable: Making an investment	(1) Agreeableness ranked first	(2) Agreeableness not ranked first
Agreeableness	-0.015	0.004
	(0.019)	(0.023)
Competitiveness	-0.046*	-0.059**
	(0.024)	(0.026)
Conscientiousness	0.014	0.021
	(0.017)	(0.018)
Extraversion	-0.014	-0.004
	(0.020)	(0.020)
Patience	0.035*	0.006
	(0.018)	(0.019)
Openness	-0.036**	-0.023
	(0.018)	(0.018)
Gender	-0.029	-0.024
	(0.021)	(0.021)
Neuroticism	-0.008	0.027
	(0.023)	(0.022)
Younger sibl.	0.016	-0.013
	(0.017)	(0.016)
Older sibl.	-0.017	0.015
	(0.018)	(0.018)
N	3420	2600
p	0.000	0.000
Adj. Rr	0.322	0.396

Table 6: Regression analyses: Confirmation bias

Notes: Regressions analyses on the subsample of treatment participants with participant and stage fixed effects. In each regression, the investment decision serves as the dependent variable. As independent variables, we include observed borrower characteristics, unobserved borrower types, a posterior dummy, and corresponding interaction terms. Column (1) shows results for treatment participants who initially ranked agreeableness as most important characteristic, while column (2) shows resuts for those who did not. Reported estimates are standardized. We cluster robust standard errors on the individual × stage level. We denote significance levels as $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.4 Treatment heterogenetities stage III

	Non-trusting		Trusting	
Dep. variable:	(1)	(2)	(3)	(4)
Making an investment	Baseline	Treatment	Baseline	Treatment
Agreeableness	0.038***	0.018**	0.056***	-0.002
	(0.012)	(0.009)	(0.009)	(0.009)
Competitiveness	-0.027**	-0.106***	-0.014*	-0.090***
	(0.012)	(0.014)	(0.008)	(0.013)
Conscientiousness	0.022**	0.022**	0.016**	0.014**
	(0.009)	(0.010)	(0.007)	(0.007)
Extraversion	0.016	0.020*	0.007	0.041***
	(0.010)	(0.010)	(0.007)	(0.008)
Patience	0.009	0.044***	-0.001	0.054***
	(0.010)	(0.010)	(0.008)	(0.010)
Openness	0.015*	0.012	0.010	0.008
	(0.009)	(0.009)	(0.007)	(0.007)
Gender (male)	-0.008	-0.038***	-0.009	-0.028***
	(0.012)	(0.010)	(0.008)	(0.010)
Neuroticism	-0.007	0.004	-0.013	0.022**
	(0.012)	(0.013)	(0.009)	(0.008)
Younger sibl. (yes)	0.005	-0.017**	0.002	-0.013**
	(0.008)	(0.008)	(0.007)	(0.007)
Older sibl. (yes)	0.011	0.018**	-0.005	0.031***
	(0.008)	(0.008)	(0.006)	(0.007)
Repayment prediction	0.150***	0.123***	0.285***	0.214***
	(0.017)	(0.014)	(0.015)	(0.016)
N	2780	2800	3340	3220
pp	0.000	0.000	0.000	0.000
Adj. R^2	0.344	0.337	0.517	0.418

Table 7: Regression analyses: Investment decisions stage III and trust

Notes: Regression analyses with participant fixed effects. In each regression, the investment decision serves as the dependent variable. As independent variables, we include observed borrower characteristics, unobserved types, and the prediction. We cluster robust standard errors on the individual × borrower level, to account for the fact that participants make one prior and one posterior decision for each borrower they encounter. We denote significance levels as $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

B Screenshots of experiment

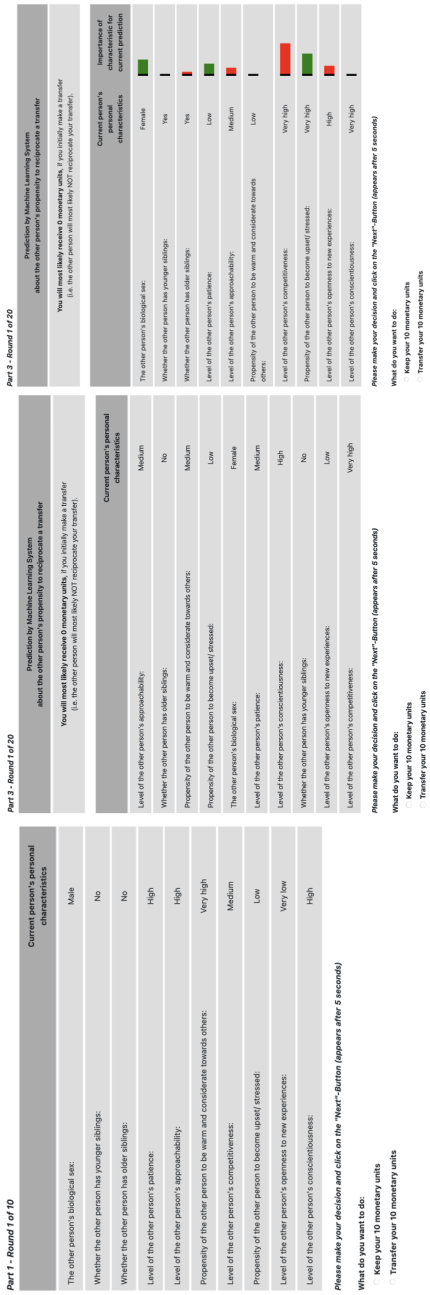

Figure 5: Interfaces of stage I, stage III—baseline, and stage III—treatment

Recent Issues

No. 314	Farshid Abdi, Mila Getmansky Sherman, Emily Kormanyos, Loriana Pelizzon, Zorka Simon	A Modern Take on Market Efficiency: The Impact of Trump's Tweets on Financial Markets
No. 313	Kevin Bauer, Andrej Gill	Mirror, Mirror on the Wall: Machine Predictions and Self-Fulfilling Prophecies
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