

Bauer, Kevin; von Zahn, Moritz; Hinz, Oliver

**Working Paper**

## Expl(AI)ned: The impact of explainable Artificial Intelligence on cognitive processes

SAFE Working Paper, No. 315

**Provided in Cooperation with:**

Leibniz Institute for Financial Research SAFE

*Suggested Citation:* Bauer, Kevin; von Zahn, Moritz; Hinz, Oliver (2022) : Expl(AI)ned: The impact of explainable Artificial Intelligence on cognitive processes, SAFE Working Paper, No. 315, Leibniz Institute for Financial Research SAFE, Frankfurt a. M.

This Version is available at:

<https://hdl.handle.net/10419/268750>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

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](mailto:info@safe-frankfurt.de) | [www.safe-frankfurt.de](http://www.safe-frankfurt.de)

# Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users’ Information Processing

Kevin Bauer

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

Moritz von Zahn, Oliver Hinz

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

Due to a growing number of initiatives and regulations, predictions of modern Artificial Intelligence (AI) systems increasingly come with explanations about why they behave the way they do. In this paper, we explore the impact of feature-based explanations on users’ information processing. We designed two complementary empirical studies where participants either made incentivized decisions on their own, with the aid of opaque predictions, or with explained predictions. In Study 1, laypeople engaged in the deliberately abstract investment game task. In Study 2, experts from the real-estate industry estimated listing prices for real German apartments. Our results indicate that the provision of feature-based explanations paves the way for AI systems to reshape users’ sense-making of information and understanding of the world around them. Specifically, explanations change users’ situational weighting of available information and evoke mental model adjustments. Crucially, mental model adjustments are subject to the confirmation bias so that misconceptions can persist and even accumulate, possibly leading to suboptimal or biased decisions. Additionally, mental model adjustments create spillover effects that alter user behavior in related yet disparate domains. Overall, this paper provides important insights into potential downstream consequences of the broad employment of modern explainable AI methods. In particular, side effects of mental model adjustments present a potential risk of manipulating user behavior and promoting discriminatory inclinations. Our findings may inform the refinement of current efforts of companies building AI systems and regulators that aim to mitigate problems associated with the black-box nature of many modern AI systems.

*Key words:* Explainable Artificial Intelligence, User Behavior, Information Processing, Mental Models

---

## 1. Introduction

Contemporary AI systems’ high predictive performance frequently comes at the expense of users’ understanding of why systems produce a certain output (Gunning et al. 2019, Meske et al. 2022). For AI systems that provide predictions to augment highly consequential processes such as hiring decisions (Hoffman et al. 2018), investment decisions (Ban et al. 2018), or medical diagnosing (Jusupow et al. 2021), this “black box” nature can create considerable downsides. These issues include impaired user trust, reduced error safeguarding, restricted contestability, and limited accountability (see Rosenfeld and Richardson 2019, for a review). Having recognized these problems, organizations developing AI and governments increasingly adopt principles and regulations (see, e.g., GoogleAI 2019, MetaAI 2021, EU 2016, 2021) effectively stipulating that AI systems need to provide meaningful explanations about why they make certain predictions (Goodman and Flaxman 2017, Cabral 2021). In light of these developments, the implementation and use of explainable AI (XAI) methods are becoming more widespread and mandated by law.

The purpose of XAI methods is to make AI systems’ hidden logic intelligible to humans by answering the question: why does an AI system make the predictions it does? Thereby, XAI methods aim to achieve high predictive performance and interpretability at the same time. Many state-of-the-art XAI techniques convey insights into AI systems’ logic post-training and explain behaviors by depicting the contribution of individual input features to the outputted prediction (Doshi-Velez and Kim 2017). While there is reason to believe that XAI can mitigate black-box problems (Bauer et al. 2021), the pivotal question is how users respond to modern explanations, given that the human factor frequently creates unanticipated, unintended consequences even in well-designed information systems (see, e.g., Willison and Warkentin 2013, Chatterjee et al. 2015).

Nascent research on human-XAI interaction examines how explainability affects humans’ perceptions, attitudes, and use of the system, e.g., trust (Erlei et al. 2020), detection of malfunctioning (Poursabzi-Sangdeh et al. 2021), (over)reliance (Bussone et al. 2015), and task performance (Senoner et al. 2021). Prior research, however, does not consider the potential consequences of providing explanations for users’ situational information processing (the use of currently available information in the given situation) and mental models (cognitive representations that encode beliefs, facts, and knowledge). By depicting the contribution of individual features to specific predictions, feature-based XAI enables users to recognize previously unknown relationships between features and ground truth labels that the AI system autonomously learned from complex data structures. In that sense, XAI may constitute the channel through which AI systems impact humans’ conceptualization and understanding of their environment. This effect could reinforce the already considerable influence contemporary AI systems have on human societies (see, e.g., Rahwan et al.

2019) by, for better or worse, allowing human users to adopt systems’ inner logic and problem-solving strategies. Despite the increasing (legally required) implementation of XAI methods, a systematic study of these effects is yet missing. The paper at hand aims to fill this important gap.

We ask three research questions: Does the additional provision of feature-based explanations affect AI system users’ situational processing of observed information? Does it affect users’ underlying mental models? What are important moderating factors? Consider, for instance, a loan officer who works with an AI system to predict an applicant’s risk parameters and determine the credit approval. Due to legal requirements (e.g., Artificial Intelligence Act (EU 2021)), the AI system recently started to provide feature-based explanations, showing that it strongly relies on people’s smartphone charging behavior to predict creditworthiness.<sup>1</sup> While previous research examines how this explanation may affect the loan officer’s perceptions of the system, we conjecture that the explanation also, and maybe more importantly, affects his processing of currently available information and his underlying mental models of the determinants of creditworthiness. By changing mental models, explanations may even reshape the loan officer’s behaviors in related domains beyond the loan approval decision, e.g., assessing the faithfulness of his daughter’s new boyfriend based on the smartphone charging behavior.<sup>2</sup>

Considerable challenges arise when trying to answer our research questions. First, measuring how XAI methods affect users’ situational processing of information and mental models is extremely difficult because these cognitive processes are typically unobserved. Second, we need to control for possible external cues, unintended stimuli, additionally attainable information, and preferences that may affect these cognitive processes in any given situation. Third, whether people interact with an (X)AI system, let alone rely on it, is highly endogenous and depends on factors such as culture, technological literacy, and the socio-technological environment. Thus, isolating effects associated with the provision of explanations in addition to predictions is particularly demanding, if not outright impracticable, in a natural (organizational) setting. To address these challenges, we rely on two complementary, incentivized experimental studies.

In Study 1 (N=607), laypeople played a series of investment games (Berg et al. 1995), making sequential economic transaction decisions in an intentionally abstract setting. In Study 2 (N=153), experts from the real-estate industry predicted listing prices for real apartments located in Germany. Study 2 extends Study 1 by testing the generalizability of our findings and elaborating on mechanisms driving the results. In both studies, conditional on the treatment, participants either

<sup>1</sup>For anecdotal evidence of such non-traditional data usage see, e.g., LenddoEFL.com or <https://money.cnn.com/2016/08/24/technology/lenddo-smartphone-battery-loan/index.html>,

<sup>2</sup>On a high level, both decisions effectively constitute sequential economic transactions under uncertainty that strongly depend on trust.

received no decision support, support from an AI system in the form of opaque predictions, or an XAI system with predictions plus feature-based explanations. We answer our research questions by eliciting and comparing changes in both participants’ decision-making patterns and their beliefs about feature-label relationships.

The two studies strongly complement each other for three reasons. First, laypeople (Study 1) and experts (Study 2) are the two diametrical archetypes of AI system users affected by growing explainability requirements. Studying both types’ responses to XAI methods enables us to identify possibly differential effects, and make inferences about the generalizability of our findings. Second, we consider two fundamental types of prediction problems where AI systems are frequently in use: transaction outcome predictions (Study 1) and price predictions (Study 2) (see, e.g., Ban et al. 2018, Rico-Juan and de La Paz 2021). Examining the two settings allows us to understand better whether the interplay between XAI and cognitive processes is task-specific. Third, employing LIME (Study 1) and SHAP explanations (Study 2) – the two most popular feature-based XAI methods (Gramegna and Giudici 2021) – allows us to draw more general conclusions about the interplay between feature-based explainability and cognitive processes.

Our findings paint a consistent picture: providing explanations is the critical factor that enables AI systems to influence the way people make sense of and leverage information, both situationally and more permanently. Crucially, we find an asymmetric enduring effect that can foster preconceptions and spill over to other decisions, thereby promoting certain (possibly biased) behaviors.

Our paper proceeds as follows. Section 2 presents theoretical foundations, while Section 3 explains our experimental studies and results. Section 4 concludes by discussing our results, the limitations of our work, and directions for future research.

## 2. Theory

In this section, we first discuss modern XAI methods (section 2.1). Subsequently, we outline the relation between providing explanations and cognitive processes (section 2.2) and discuss our work’s contribution to the literature (section 2.3).

### 2.1. Explainable Artificial Intelligence

Following Doshi-Velez and Kim (2017), we conceptualize XAI as methods that possess the ability to present in understandable terms to a human why an AI system makes certain predictions. Over the last couple of years, researchers developed ample XAI methods that help elucidate the opaque logic of machine learning (ML) based AI systems (see, e.g., Ribeiro et al. 2016, Lundberg and Lee 2017, Koh and Liang 2017, Lakkaraju et al. 2019). Very generally, XAI methods aim to alleviate problems associated with the black-box nature (e.g., distrust, lack of accountability, and error safeguarding) while maintaining a high level of prediction accuracy (Bauer et al. 2021).

Our study focuses on feature-based XAI methods – hereafter XAI methods – that can explain the behavior of any ML-based AI system by showing the contribution of individual features to the prediction. We do so for several reasons. First, these explanations are the most widespread in practice (Bhatt et al. 2020, Senoner et al. 2021, Gramegna and Giudici 2021). Second, they are highly intuitive and straightforward to interpret as they satisfy most requirements for human-friendly explanations (Molnar 2020). Third, they are typically applicable to systems using structured and unstructured data (see, e.g., Garreau and Luxburg 2020). Fourth, these methods can explain individual predictions – local explainability – which might be the only method legally compliant with (upcoming) regulations (Goodman and Flaxman 2017).

Many researchers recognize two related XAI methods as state-of-the-art: LIME and SHAP (Gramegna and Giudici 2021, Molnar 2020). LIME (Ribeiro et al. 2016) and SHAP (Lundberg and Lee 2017) provide explanations through additive feature attributions, i.e., linear models that depict the numeric contribution of each feature value to the overall black box model prediction. Both approaches learn these interpretable “surrogate models” on input-prediction pairs of the black box model and are applicable to virtually all classes of ML models, i.e., are model agnostic. On the individual level, SHAP and LIME provide contrastive explanations that inform users why predictions for a specific instance diverge from the prediction for an average instance (Molnar 2020). For example, if the SHAP value for the feature *Balcony* equals +500 (-200), it indicates that having a balcony marginally increases (decreases) the current apartment’s listing price prediction by 500\$ (200\$). The big difference between LIME and SHAP is the way of estimating the additive feature attributions. LIME creates synthetic, perturbed data points in the local neighborhood of the observation of interest and fits a weighted linear model to explain the relationship between the synthetic data and the relevant black box predictions. Importantly, LIME weights synthetic instances based on their proximity to the original data point. By contrast, SHAP is inspired by coalitional game theory and treats input features as a team of players that cooperate to generate a payoff (the prediction). The method essentially estimates the marginal contribution of each player to the overall payoff – Shapley values (Shapley 1953) – using a linear model that weights instances based on characteristics of coalitions. Given these mathematical differences, the two methods can produce (slightly) different feature-attributions for the same instance. However, from the perspective of a user who is not familiar with these details, the intuition and interpretation of the two methods’ explanations are reasonably similar (Molnar 2020). Notably, LIME and SHAP closely relate to Gregor and Benbasat’s (1999) seminal description of “why and why not explanations” in the context of knowledge-based expert systems.

With the development of modern explainability methods, research on the impact of contemporary XAI on user behavior has become increasingly essential (Vilone and Longo 2021). Nascent

research in this domain typically focuses on how explanations affect user attitudes and reliance on the AI system (see, e.g., Lu and Yin 2021). These studies produce mixed evidence on the consequences of XAI on decision performance, user trust, perception, and decision-making performance. Several studies depict that explanations can enhance trust in and positive perceptions of the system (see, e.g., Dodge et al. 2019, Rader et al. 2018, Yang et al. 2020), whereas others provide reversed evidence (see, e.g., Erlei et al. 2020, Poursabzi-Sangdeh et al. 2021). While prior studies produce important insights regarding the interplay between XAI and user perceptions, none of them considers that the additional provision of explanations may also reshape users’ information processing, both situationally and more permanently. For instance, employing SHAP to show the contribution of input features to a creditworthiness prediction may not only affect a loan officer’s perception of the AI system in use. Instead, she may process currently available information about the applicant differently and develop a novel understanding of the determinants of creditworthiness, i.e., adjust her mental model. With the increasing adoption of explainability principles by organizations (see, e.g., GoogleAI 2019, MetaAI 2021) and the growing number of regulatory transparency requirements (see, e.g., EU 2016, 2021), it is pivotal to understand how contemporary XAI methods influence cognitive processes that lie at the heart of people’s knowledge, behavior, and problem-solving capabilities.

## 2.2. Cognitive Perspective on XAI Employment

Through feature-based explanations about an AI system’s prediction, human users can observe possibly unknown feature-label relationships that the system learned from complex data structures by itself (Agarwal and Dhar 2014, Berente et al. 2021). While providing explanations, in general, can have a variety of cognitive effects, researchers across disciplines generally agree that they primarily enhance people’s understanding of someone or something, improve reasoning, and facilitate learning (Malle 2006, Gregor 2006). From a cognitive perspective, obtaining explanations can entail two effects: First, it may change people’s situational processing of available information – their use of available information while observing explanations. Second, it can lead to an adjustment of their beliefs about feature-label relationships the AI system inherently models – their mental representation of real-world processes. In this paper, we follow previous work in information systems and rely on the “Mental Models Framework” to conceptualize relevant cognitive processes (see, e.g., Vandenbosch and Higgins 1996, Lim et al. 1997, Alavi et al. 2002).

Mental models are “*all forms of mental representation, general or specific, from any domain, causal, intentional or spatial*” (Brewer 1987, p.193), encoding beliefs, facts, and knowledge (Jones et al. 2011). Through imaginary manipulations of model components, people can reason and make inferences about how to solve problems (Rouse and Morris 1986). Much of the people’s decision-making is based on these simulations which figuratively create informal algorithms for carrying out



specific tasks (see, e.g., Johnson-Laird et al. 2017). For instance, real estate agents can mentally simulate how listing prices might change if an apartment for sale had a balcony.

When people perform tasks, they draw upon relevant mental models that guide their processing of incoming information to form expectations and make (expectedly) optimal decisions. Working with an AI system that provides black box predictions, i.e., information relevant to the task, allows people to reflect on their own expectations and compare it to the machine prediction (Schön 2017). This mental process might entice people to revise their expectations and thus make different decisions because the machine prediction effectively substitutes for people’s own mental model driven formation of expectations (Agrawal et al. 2019). However, the black box nature does not allow users to directly compare their underlying beliefs and logic with that of the AI system. This comparison can only occur when they learn how the system combines available information to arrive at a prediction. In the previous example, the real estate agent may have access to an XAI system that provides a listing price prediction together with an explanation of how specific apartment attributes contribute to it. The agent can compare the explanation to her own initial perception of the individual attribute contributions to the listing price. As a result, the agent may detect inconsistencies that prompt her to revise her logic by putting more or less emphasis on specific information currently available to evaluate the apartment. This explanation-enabled situational process (Schön 2017) can reconcile the distinct logic that humans and machines apply to arrive at a certain assessment. From this perspective, providing explanations on top of predictions may constitute a pivotal factor in allowing users to reflect on how they leverage information to solve a problem and adapt it according to the AI system’s logic for the given task.

Apart from situationally changing cognitive processes that shape the current decision, the interaction between mental models and explanations may also yield lasting effects because mental models possess the dynamic capacity to change (Jones et al. 2011). Repeatedly observing explanations about how feature  $X$  contributes to prediction  $\hat{Y}$  and engaging in reflection processes may evoke adjustments of the underlying mental model in use. Following Vandenbosch and Higgins (1996), exposure to external stimuli – here explanations – can lead to two mental model adjustment processes: maintenance and building. Under mental model maintenance, people feel encouraged to maintain or reinforce current beliefs and decision-making rules. This process occurs when they perceive or select new information to fit into their current beliefs and routines. Under mental model building, individuals profoundly restructure or build new mental models in response to handling novel, disconfirming information. As a result of these processes, individuals may adopt different beliefs about how  $X$  contributes to the real label  $Y$ , enticing them to process information differently even when explanations are no longer present. Put differently: users may not merely combine situationally observed explanations with their own logic to solve a given task. Instead, observing

the system’s logic may more fundamentally reshape users’ way of solving problems in general, i.e., evoke learning. Therefore, users may exhibit different problem-solving strategies whenever they draw upon the explanation-adjusted mental model, even in situations where they do not observe explanations anymore.

In sum, cognitive theories give reason to believe that providing explanations in addition to predictions can influence users’ processing of information about feature  $X$ , both situationally and more fundamentally. Due to the latter effect, modern XAI methods may constitute a cornerstone of effective knowledge transfers from ML-based AI systems to human users, helping them to learn from the AI how  $X$  relates to  $Y$ . Hence, explanations could facilitate learning *machine knowledge* – new knowledge AI systems autonomously learned from Big Data and previously missed by domain experts (Teodorescu et al. 2021, van den Broek et al. 2021).

### 2.3. Contribution to the Literature

Our study complements three different streams of literature. The first and most closely related line of work studies the interplay between XAI techniques and user behavior (see Rosenfeld and Richardson 2019, Vilone and Longo 2021, for an overview). About two decades ago, several studies found that suitably designed explanations about the functioning and purpose of legacy knowledge-based expert systems can increase users’ trust in the systems, improve users’ perceptions of the system, and enhance decision-making performance (Dhaliwal and Benbasat 1996, Gregor and Benbasat 1999, Ji-Ye Mao 2000, Wang and Benbasat 2007). However, these expert systems codify knowledge from human experts as explicit procedures, instructions, rules, and constraints in a digital format. They do not represent machine knowledge that modern ML-based AI systems learn independently of domain experts by training on large data sets (van den Broek et al. 2021). Given the inherent distinctions between expert systems and ML-based AI systems in terms of encoded knowledge, contemporary explainability methods present an entirely different form of reasoning to users, namely that of machines (Vilone and Longo 2021, Meske et al. 2022). More recent research on the impact of explainability on user behavior mainly focuses on how contemporary XAI methods impact users’ perceptions of the AI system. This nascent literature shows that explainability often improves reliance on and trust in the system (Bussone et al. 2015), fairness perceptions (Dodge et al. 2019), human-AI collaboration (Yang et al. 2020), task efficiency (Senoner et al. 2021), and users’ understanding of the system’s malfunctions (Rader et al. 2018). However, there is also evidence of disadvantages relating to informational overload (Poursabzi-Sangdeh et al. 2021), reduced user trust (Erlei et al. 2020), and overreliance (Bussone et al. 2015). Moreover, explanations that are unstable and sensitive even to small perturbations to inputs have the potential to mislead human users into trusting a problematic black box, e.g., by selectively providing explanations that conceal biased behaviors and malfunctions (Lakkaraju and Bastani 2020, Kaur et al.

2020). Hence, explanations may be a security concern if adversaries use perturbations of inputs and model attributes to produce intentionally misleading explanations that manipulate users' trust and behaviors (Ghorbani et al. 2019). We complement this pivotal and insightful work by examining the impact of contemporary XAI on users' situational information processing and mental models. Understanding how the provision of explanations about the workings of ML-based AI systems may reshape these cognitive processes is pivotal for anticipating the downstream consequences of this technology on human societies and designing effective transparency and explainability regulations.

The second literature we complement explores the mechanisms of learning in socio-technological environments. A common theoretical foundation builds upon Bayes rule as a rational benchmark of how humans accommodate new information (see, e.g., Holt and Smith 2009). However, research has shown systematic deviations from Bayes' rule. Reasons include over- or underweighting of new information (Rabin and Schrag 1999) and a general tendency to asymmetrically discount information conflicting with prior beliefs while readily internalizing confirming information (Yin et al. 2016). We complement this research stream by showing how human users deviate from Bayes rule in the context of learning from modern AI systems. Notably, there exists a limited number of prior research examining how black box predictions change users' decision-making habits (Abdel-Karim et al. 2020, 2022, Jussupow et al. 2021, Fügner et al. 2021a,b). Relatedly, in a formal model, Agrawal et al. (2019) show that the predictions of black box AI systems can alter users' abilities by providing them with incentives to learn to assess the (negative) consequences of their actions for the task supported by the AI.<sup>3</sup> None of these studies, however, examines the role of feature-based explanations in learning, which could pave the way for more fundamental changes in the way users understand real-world processes. Our paper intends to fill this gap. We study how the provision of explanations about how an AI system solves prediction tasks allows users to integrate the presented machine knowledge into their mental models, i.e., learn from XAI. A better understanding of how explainability may contribute to *machine teaching* – the notion that AI systems first learn novel knowledge that experts neither conceive nor anticipate from data and then transfer this knowledge to human users (Abdel-Karim et al. 2020) – is particularly significant given the growing requirements to implement explainability methods when using AI systems.

The third stream of literature we add to studies how humans collaborate with computerized systems to solve problems. Previous research in this area dates back decades. Several studies document

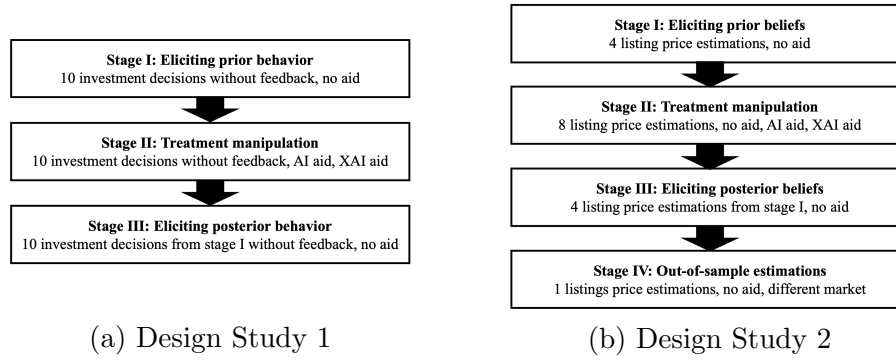
<sup>3</sup> Explainability may enter Agrawal et al.'s model by changing the prediction reliability. Following proposition 2, the necessity for providing explanations decreases with the users' judgment. However, the model does not consider the idea presented in our paper that explainability may also affect users' understanding of the process that determines the uncertain state of the world the AI tries to predict. One could integrate this notion into the framework by modeling that explanations affect users' judgment capabilities by influencing beliefs about underlying processes. Extending Agrawal et al.'s model in this direction may be a fruitful endeavor to understand better whether explainability modulates the relationship between prediction and judgment. However, an extension of the formal model is beyond the scope of this paper and left for future research.

that humans resist using computerized decision aids, despite possible performance benefits (e.g., Kleinmuntz 1990), while others find that humans possess a strong preference for using them (e.g., Dijkstra 1999). With the growing employment of modern AI systems in a broad range of domains, the examination of human-machine collaboration has seen a considerable resurgence, e.g., in the domain of finance (Ge et al. 2021), medicine (Jussupow et al. 2021), customer service (Schanke et al. 2021), and on-demand tasks (Fügener et al. 2021a). Research on “centaur” systems (e.g., Goldstein et al. 2017, Case 2018) documents how hybrid human-AI systems (i.e., centaur systems) achieve superior results in comparison to the entities operating independently (see, e.g., Dellermann et al. 2019, Tschandl et al. 2020), promising considerable benefits from successful human-AI collaboration. Several factors moderate the interaction of humans and AI systems including the perceived subjectivity of the task (Castelo et al. 2019, Logg et al. 2019), seeing the system err (Dietvorst et al. 2015), being able to modify predictions (Dietvorst et al. 2018), the divergence between actual and expected predictive performance (Jussupow et al. 2020), and, most importantly for our research, understanding the system’s internal logic (Gregor and Benbasat 1999, Hemmer et al. 2021). Following our conjecture that explanations pave the way for AI systems to affect people’s cognitive processes, contemporary XAI methods introduce another layer of complexity in human-AI interaction and its success: an interaction between machine and human problem solving strategies. Our work provides novel insights into whether and under what circumstances people prefer to rely on their own way of leveraging information or willingly adjust it according to machine explanations. In this sense, our work contributes to the literature on (hybrid) human-AI collaboration by analyzing the underlying cognitive processes that may facilitate or hinder the realization of the promise of this technology.

### 3. Empirical Studies

We now present the design and results of Studies 1 and 2. In both studies, participants made decisions under uncertainty (providing loans and predicting apartment listing prices) either with the aid of an opaque AI, an explainable AI or without any support. We paid participants according to their decision-making performance to reveal actual preferences and beliefs.<sup>4</sup> We implemented both studies using oTree, Python, and HTML and ran them online. In Study 1, we recruited 607 participants on Prolific and let them engage in deliberately abstract investment games (Berg et al. 1995). Results allow us to observe how the provision of explanations on top of predictions shapes information processing and mental models for laypeople in a very general sequential transaction domain. Study 2 extends the first study by testing the generalizability of mental model adjustments

<sup>4</sup> See the supplementary material for details on the experimental procedures including payments, instructions, and screenshots.



**Figure 1** Structure of empirical studies

Notes: We provide an overview of the main sequence of our two empirical studies. Panel (a) and (b) respectively show how Studies 1 and 2 proceed.

regarding the task domain (listing price predictions), decision-maker expertise, and the explanation presentation, and elaborates on important asymmetric effects. With the help of our industry partner the *Real Estate Association Germany (IVD)* we recruited 153 experts from the real estate industry to participate in Study 2. We report the designs and results of the two studies consecutively. Figure 1 portrays an overview of the experimental designs.

### 3.1. Study 1

**3.1.1. Design.** In Study 1, participants repeatedly engaged in one-shot investment games (Berg et al. 1995) that possess the following structure. An investor receives 10 monetary units (MU). The investor initially observes ten deliberately abstract borrower characteristics and decides whether or not to invest her 10 MU with the borrower. If she does not invest, the game ends without the borrower making a decision and both the investor and borrower earn a payoff of 10 MU. If she invests, the borrower possesses 30 MU and can keep the whole amount without repercussions. Crucially, the borrower can repay the investor 10 MU, thereby reciprocating the investor’s initial trust. In case of repayment, the investor receives 20 MU (we double the amount); otherwise, the investor earns 0 MU while the borrower gets 30 MU. The borrower, in the absence of sufficiently strong social motives, e.g., altruism, egalitarian concerns, or moral preferences (see, e.g., Miettinen et al. 2020), will not make a repayment and maximize his personal income. As a result, the payoff structure of the investment game is of an adversarial nature from the investor’s perspective since her material well-being is at the mercy of the borrower if she invests. The investor loses her initial investment of 10 MU whenever the borrower pursues pure income-maximizing or adversarial motives like wanting to minimize the investors’ payoffs. Given this payoff structure, an income-maximizing investor in the experiment will only invest if (i) her belief that the borrower’s motive leads him to repay her is sufficiently strong, and (ii) she ultimately judges that the prospect of

doubling her income is worth risking the loss of her investment.<sup>5</sup> Study 1 participants always played as investors. Borrowers are subjects from a previous incentivized field study who had to decide upon repayment assuming an initial investment, i.e., they have already committed to a repayment decision and cannot strategically change this choice ex-post. We did not provide intermediary feedback to prevent the development of idiosyncratic expertise, experience, or investment strategies that may confound our results. We randomly matched investor and borrower decisions to determine game outcomes at the end of the study and pay both according to the earned MU.

Study 1 comprised a baseline (AI) and a treatment (XAI) condition, each with three stages.<sup>6</sup> In Stage I, each participant made 10 investment decisions for distinct, randomly drawn borrowers without intermediary feedback. They always observed the ten characteristics of a borrower and did not obtain any aid. The idea is that the ten borrower characteristics allow investors to get an idea of the likelihood that an individual borrower will make a repayment – for whatever motives – and to assess whether it is worth taking the risk of losing their investment. We deliberately chose ten unintuitive traits correlated with a person’s repayment inclination so that participants did not possess strong prior beliefs about the informativeness of characteristics for someone’s repayment behavior (see Table 4 in the supplementary material).

Stage II introduced our treatment variation. Participants made 20 decisions for new random borrowers observing all ten borrower traits. Additionally, baseline participants saw an AI system’s prediction about whether borrowers will repay an initial investment. Again we did not provide intermediary feedback. We trained the AI system on 1,054 distinct data points collected in a previous field study, the same data set that the borrowers that participants encounter in the experiment stem from (see the supplementary material for details). The system did not continue to learn during the experiment. Treatment participants, on top of predictions, observed LIME explanations (Ribeiro et al. 2016) for each borrower characteristic, informing them of its contribution to the repayment prediction. Revealing LIME values on top of identical predictions constituted the treatment variation. As is often the case, we depicted LIME values graphically using colored bars of different lengths. Participants received detailed information about the model, input features, performance on a representative test set, and how to interpret LIME explanations.

<sup>5</sup> When a risk-neutral, purely self-interested investor expects that the borrower repays her with a probability of  $p > 0.5$ , e.g., because she believes the borrower to possess altruistic, efficiency, or fairness preferences, they have a strict incentive to invest because they maximize their expected earning. Importantly, holding such expectations about the borrower’s preferences is justified and frequently observed in sequential games – a considerable share of people does respond reciprocally in sequential exchanges if they are trusted (see, e.g., Miettinen et al. 2020, for an overview).

<sup>6</sup> Note: To reduce the complexity for the reader, we only report the three main stages of the experiment. Right before and after Stage II, we additionally measured participants’ prior and posterior preferences to observe three borrower characteristics. We use these measures as robustness and consistency checks. We provide a detailed description of these measurements in the supplementary material.

Stage III perfectly mirrored Stage I. Importantly, participants engaged with the same borrowers from Stage I in random order. We did not draw participants’ attention to this fact to alleviate concerns about the experimenter’s demand effect. The study concluded with a brief questionnaire on socio-economic control variables.

**3.1.2. Results.** Throughout our analyses of Study 1, we mainly rely on the following regression model:

$$Y_{ijs} = \beta_1 \cdot X_j + \beta_2 \cdot (X_j \times I_s) + \beta_3 \cdot (X_j \times Expl_i) + \beta_4 \cdot (X_j \times Expl_i \times I_s) + \gamma_{ijs} + \epsilon. \quad (1)$$

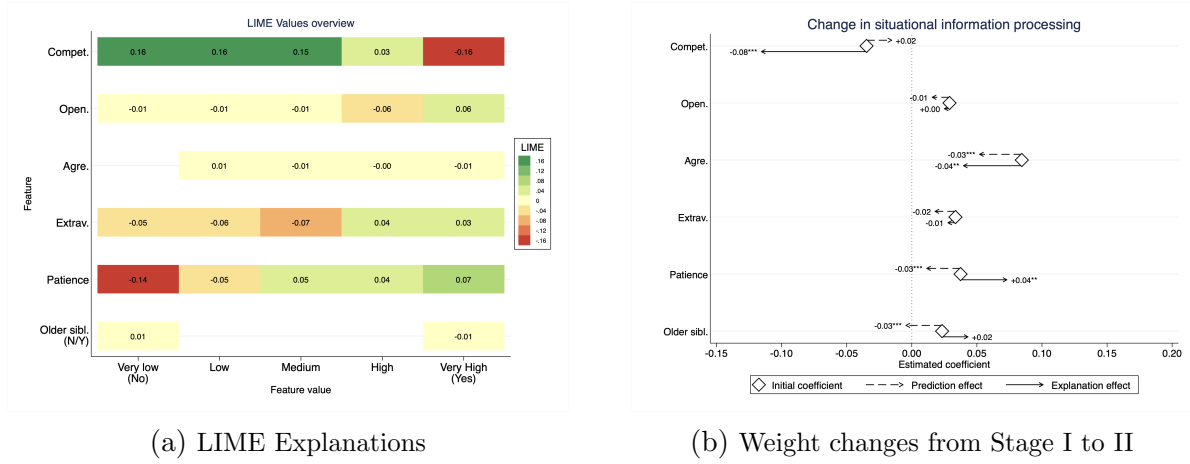
$Y_{ijs}$  is a dummy indicating whether participant  $i$  invested with borrower  $j$  in Stage  $s$ . Hence,  $\beta$  coefficients measure variation in the probability to invest with a borrower.  $X_j$  is a vector reflecting the ten observed borrower traits, the overall prediction, and LIME values.<sup>7</sup> Most relevant to our analyses,  $I_s$  and  $Expl_i$  are dummy variables respectively indicating whether a decision takes place in Stage  $s$  compared to Stage I (i.e., Stage I serves as the reference category) and whether participant  $i$  is in the XAI treatment (observes explanations on top of predictions in Stage II).  $\gamma_{is}$  represents individual-state fixed effects. We report standardized regression coefficients with robust standard errors. Our main interest lies in the interaction terms  $\beta_3$  and  $\beta_4$  respectively capturing the isolated effects of observing the prediction and additionally observing LIME explanations. As  $\beta_4$  constitutes a Difference-in-Difference (DiD) estimator, it is pivotal to check that before the intervention, there are no treatment differences (parallel trends assumption). Regression analyses reveal that baseline and treatment participants in Stage I did not place significantly different weight on any trait, hence the use of a DiD identification strategy appears generally valid. Nevertheless, because participants placed significant weight on *Gender*, *Conscientiousness*, *Neuroticism*, and *Younger Siblings* in only one of the two conditions participants, there is still some concern about the appropriate interpretation of DiD estimates for these traits.<sup>8</sup> To avoid drawing incorrect conclusions, we conservatively refrain from interpreting these traits’ estimates.

**Situational information processing.** We start analyzing how participants’ weighting of borrower characteristics changed from Stage I to II, i.e., changes in participants’ situational information processing. Figure 2 illustrates our results. Panel (a) depicts the average LIME values (color saturation) participants observed for different feature values (y- and x-axis). Higher positive (negative) LIME values depict a higher positive (negative) contribution of a given feature value to the

<sup>7</sup> For most traits, values and LIME values are almost perfectly correlated producing severe problems of multicollinearity (see Table 7 in the supplementary material). Therefore, in our regression analyses, we only include LIME explanations for which there exists a tolerable correlation between the trait and LIME values: Openness, Agreeableness, and Conscientiousness.

<sup>8</sup> See Table 8 in the supplementary material.

predicted probability that a borrower makes a repayment. Panel (b) portrays how the provision of predictions and explanations affected the weighting of a given borrower trait. The diamond marker represents the original weighting in Stage I ( $\beta_1$ ). The dashed and solid arrows respectively illustrate the isolated effects of observing predictions ( $\beta_3$ ) and additional explanations ( $\beta_4$ ). Depicted results stem from regressions reported in Table 9 in the supplementary material.



**Figure 2** Illustration of prediction and explanation effects on situational information processing.

Notes: We illustrate how the provision of opaque predictions and LIME explanations on top of predictions affect participants' situational information processing. Panel (a) shows the LIME values (z-axis) for different feature values (x-axis) participants observed in the study. For the binary feature Older siblings, we show the LIME values for No and Yes at the outer limits of the continuous feature scale. Panel (b) depicts with the estimated prediction and explanation effects – respectively  $\beta_3$  and  $\beta_4$  in model (1) with  $s = 2$ . Initial values represent  $\beta_1$ . We denote significance levels by \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

There are two main insights. First, prediction effects in Panel (b) suggest that the provision of opaque predictions generally decreased the weight participants placed on observed borrower traits. Although only the estimates for *Agreeableness*, *Patience*, and *Older Siblings* are significant, predictions reduced the absolute magnitude of all variables. Second, the provision of explanations on top of predictions entailed significant weight changes that mirror the relationship between borrower traits and repayment behavior as depicted by the LIME values. Panel (a) shows that the predicted repayment probability markedly decreases (increases) with a borrower's level of *Competitiveness* (*Patience*). Panel (b) reveals that these are the two traits whose weighting the provision of explanations significantly fostered: observing explanations rendered the relationship between a borrower's *Competitiveness* (*Patience*) and a participant's investment likelihood significantly more negative (positive). LIME values reveal that *Agreeableness* – the trait participants initially weighted the most – has almost no impact on the repayment prediction. Accordingly, we find that the provision of explanations led to a significant decrease in the magnitude of the weight participants placed on this trait. Additional analyses confirm that LIME values for these three characteristics had a significantly positive influence on participants' investment decisions, corroborating the notion that



participants paid attention to and adjusted their weighting of traits according to observed explanations (see Table 11 in the supplementary material). Taken together, participants significantly adjusted their weighting of information in the direction of observed explanations for (i) the trait they initially perceived as most important, and (ii) the traits LIME highlighted as most important.<sup>9</sup> Finally, while not shown in the Figure 2 for ease of interpretation, regression analyses further reveal that explanations significantly reduced the weight participants placed on the prediction as such, i.e., they were less likely to follow a prediction that a borrower makes a repayment.<sup>10</sup>

**Result 1.1:** *Observing explanations changed participants’ situational processing of the overall prediction and borrower traits that explanations or they themselves consider most important. The direction of adjustments mirrors explanations.*

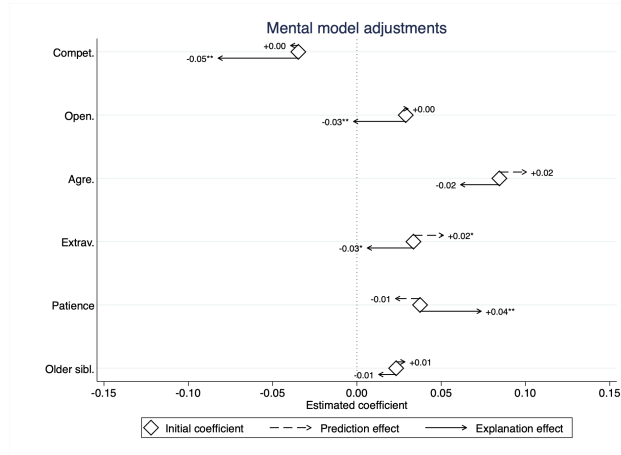
Result 1.1 accords with our theoretical elaborations: people adjust their situational information processing in response and according to explanations they currently observe. Notably, elicited expectations about the prediction accuracy did not differ significantly for predictions with or without explanations (71.8% and 70.6% respectively,  $p = 0.751$ , Wilcoxon rank-sum test). Therefore, changes in the weighting of predictions do not seem to result from lower performance expectations. Next, we test the conjecture that explanations affect beliefs about the relationship between borrower characteristics and repayment behavior, i.e., mental models.

**Mental model adjustments.** We compare participants’ information weighting across Stages I and III, to test the conjecture that explanations affect mental models about the relationship between borrower traits and repayment behavior. We rely on the regression model (1), setting  $s = 3$  and excluding controls for the prediction and LIME values. Figure 3 illustrates regression results which we report in Table 12 in the supplementary material.

Figure 3 portrays how the provision of predictions and explanations lastingly changed the weighting of a given borrower trait across Stages I and III, where participants had no (X)AI aid. The

<sup>9</sup> Note that these results do not allow us to isolate how explanations affect what investors consider to be a borrower’s motivation to repay them or not. The change in the weighting of competitiveness could stem from a reinforced perception that competitiveness predicts a low repayment likelihood because it proxies for anti-social, income-maximizing, or relative income-maximizing motives. While we cannot isolate investors’ latent belief(s) about borrowers’ motives, our results effectively show that the provision of explanations does entail a change in at least one of these perceived latent motives, i.e., that XAI can change the processing of information. A similar argument applies regarding mental model adjustments outlined below.

<sup>10</sup> Reported results are robust to excluding participants who always or never invested in our analyses, respectively alleviating concerns that our results are driven by pure altruists or players who always choose the game-theoretically dominant strategy (see the subsection on additional robustness checks in the supplementary material). Instead, our results stem from those participants whose behavior suggests that they try to invest with borrowers whom they believe will make a repayment, i.e., individuals who, from a conceptual point of view, should be most inclined to learn to recognize repaying borrowers. Results 1.2 and 1.3 are equally robust to excluding these “extreme” types, warranting a similar interpretation.



**Figure 3** Mental model adjustments.

Notes: We depict participants' mental model adjustments as measured by their change in the weighting of borrower traits across Stages I and III. The estimated prediction and explanation effects respectively represent  $\beta_3$  and  $\beta_4$  in model (1) with  $s = 3$ . Initial values represent  $\beta_1$ . We denote significance levels by \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

diamond marker depicts the original weighting in Stage I ( $\beta_1$ ). The dashed and solid arrows respectively show how having observed predictions ( $\beta_3$ ) and explanations on top of predictions ( $\beta_4$ ) did fundamentally alter participants' information processing, i.e., mental models.

Observing opaque predictions did not result in a significant change in participants' weighting of borrower traits. By contrast, depicted results suggest that providing explanations entailed an asymmetric adjustment of mental models. Specifically, explanations led participants to place significantly more weight on borrowers' *Competitiveness* and *Patience* in Stage III than in Stage I. The weight changes again mirror the observed LIME explanations. After observing explanations that the AI system places the most weight on borrowers' *Competitiveness* and *Patience*, participants increased their weighting of these attributes even for investment decisions where they no longer observed explanations. Intriguingly, we do not find that explanations about the low relevance of *Agreeableness* led participants to adjust their marked weighting of this trait significantly. Although participants weighted *Agreeableness* significantly less while observing explanations, they returned to their original weighting of it once they lost access to the XAI system. Naturally, one may wonder about this asymmetry's origins. One plausible interpretation is that explanations are less likely to evoke pronounced mental model adjustments when they conflict with strong preconceptions. Put differently: people are more inclined to engage in mental model maintenance rather than building because it is less cognitively demanding and creates less psychological distress (Vandenbosch and Higgins 1996). In Stage I, participants put by far the most emphasis on a borrower's *Agreeableness* to decide upon investing. LIME values, however, suggested that this conception is incorrect because it is among the least relevant predictors for borrowers' repayment inclination. Even though one would expect that participants engaged in mental model building to reshape their beliefs about the

relationship between *Agreeableness* and repayment behavior, we do not find significant adjustments. For *Competitiveness* (*Patience*), explanations depicted an important negative (positive) influence which, given their initial weighting of it, confirmed participants’ prior beliefs. Following the Mental Models framework, confirming explanations should evoke the maintenance or reinforcement of prior beliefs. Given the significant explanation effects, it seems that participants willingly engaged in this process. This inclination to engage in mental model maintenance rather than building more generally concurs with the frequently documented confirmation bias (see, e.g., Yin et al. 2016), i.e., the tendency to selectively process information in a way that allows for the continuation or strengthening of beliefs. We elaborate on this issue in Study 2 and the discussion.<sup>11</sup>

**Result 1.2:** *Machine explanations entailed asymmetric mental model adjustments. Participants reinforced priors that explanations confirmed but did not abandon priors that explanations markedly contradicted.*

**Investment performance.** So far it remains open how providing explanations on top of predictions affected participants’ decision-making performance in our setting. Table 1 summarizes participants’ performance measured by the accuracy (share of payoff maximizing decisions) and recall (share of investments with repaying borrowers). We also report  $p$ -values of  $F$ -tests to illustrate significant treatment differences.<sup>12</sup>

	Stage I (no aid)		Stage II (with aid)		Stage III (no aid)	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
Baseline (AI) in %	60.3	64.9	63.1	64.6	62.7	65.1
Treatment (XAI) in %	60.7	67.4	57.5	57.5	56.5	60.2
F-test: Base. v. Treat.	$p = 0.79$	$p = 0.31$	$p < 0.01^{***}$	$p < 0.01^{***}$	$p < 0.02^{**}$	$p < 0.04^{**}$

**Table 1** Investment performance across stages.

Notes: We depict participants’ investment performance as measured by their accuracy (share of payoff maximizing decisions) and recall (share of investments with repaying borrowers) in Stages I, II, and III. We report results separately for Baseline (AI) and Treatment (XAI) participants.  $F$ -tests reveal the significance of treatment differences per measure and stage. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

While there are no differences in Stage I, treatment participants performed significantly worse than baseline ones in Stage II.<sup>13</sup> Treatment participants’ relatively lower performance in Stage II

<sup>11</sup> Note: the significant explanation effect for *Openness* and *Extraversion* may be a consequence of participants’ significantly stronger weighting of borrowers’ *Competitiveness* and *Patience* and a limited capacity to process information. Specifically, XAI participants in Stage III place similarly low weight on all borrower traits but *Competitiveness*, *Agreeableness*, and *Patience*. This pattern may suggest that participants heuristically focus on the three characteristics that they themselves and the AI system deemed most relevant to the decision. As a result, they place less weight on all other traits, which for *Openness* led to a statistically significant effect.

<sup>12</sup> We show ROC curves in Figures 17 to 19 in the supplementary material

<sup>13</sup> Note: Participants neither knew their own nor the AI system’s performance because we did not provide intermediate feedback. Therefore, they could not see how much better or worse the system performs compared to themselves. While unknown to participants, predictions are accurate in about 69.3% of the cases. This performance holds equally

stems from not investing with the most competitive borrowers (with most negative LIME values) while the overall prediction implies doing so, i.e., from overruling positive predictions.<sup>14</sup>

They overruled positive predictions and refrained from investing in 46.5% of these cases, resulting in a decision accuracy of merely 53.5%. Baseline participants, for most competitive borrowers, overruled positive predictions only in 21.2% of the cases and achieved a decision accuracy of 78.9%. For all other borrowers, treatment (baseline) participants overruled positive predictions and made optimal decisions in 23% (19.4%) and 69.6% (71.1%) of the cases, respectively. Hence, treatment participants seem to have placed too much weight on very high competitiveness, leading them to overrule the overall prediction inefficiently often.

Examining Stage III, we find that this overweighting of the highest competitiveness level persisted even when participants did not observe explanations anymore (see Table 13). In Stage III, treatment (baseline) participants invested with most competitive borrowers in 44.7% (54.7%,  $p < 0.01$ ,  $F$ -test) of the cases; with other borrowers in 68.2% (67.6%,  $p = 0.7$ ,  $F$ -test) of the cases. As a result, treatment (baseline) participants achieved a decision accuracy of 51.7% (57.2%,  $p < 0.01$ ,  $F$ -test) for most competitive borrowers and 59.5% (62.8%,  $p < 0.05$ ,  $F$ -test) for other borrowers. Notably, participants already associated very high competitiveness with a low repayment likelihood in Stage I: most competitive borrowers received an investment in 56.3% of the cases, while all others did so in 69.5% of the cases (there do not exist treatment differences). Against this background, explanations seem to have exacerbated this inaccurate pattern<sup>15</sup> to an extent that treatment participants made significantly worse decisions than before. Put differently, confirming explanations inappropriately reinforced preconceptions about most competitive borrowers not repaying an investment in our setting.

**Result 1.3:** *Participants excessively increased the isolated weighting of a trait they already believe to be evidence against repayment. This reaction inefficiently decreased participants' likelihood to invest with repaying borrowers that were highly competitive.*

for both repaying (69.7%) and non-repaying borrowers (67.7%). Participants in Stage I correctly invest with (non-)repaying borrowers in 66.1% (41.2%) of the cases and overall in 60.5% of the cases. Put differently, the AI system outperforms them regarding both types of borrowers and especially for the identification of non-repaying ones. As a result, participants could have benefited from relying on the predictions – which baseline participants did at least partially.

<sup>14</sup> Across Stages I and II, baseline participants' access to the AI system significantly increased the accuracy by 4.6% ( $p < 0.01$ ,  $F$ -test), whereas the recall effectively remained constant ( $p = 0.82$ ,  $F$ -test). XAI participants' performance significantly decreased regarding both the accuracy (-5.3%;  $p < 0.01$ ,  $F$ -test) and recall score (-14.6%;  $p < 0.01$ ,  $F$ -test).

<sup>15</sup> A purely linear distinction between most competitive and other borrowers does not allow to draw conclusions about their repayment likelihood: they respectively made a repayment in 77.4% and 79.8% of the cases ( $p = 0.85$ ,  $F$ -test).

In sum, the results for Study 1 are highly consistent with the notion that the provision of explanations creates a novel channel through which AI systems may reshape users’ way of processing information, both situationally and more permanently. For the latter effect, we observe an asymmetry that is reminiscent of a confirmation bias and, in our setting, decreased participants’ decision-making performance by excessively reinforcing inaccurate preconceptions.

### 3.2. Study 2

The goal of Study 2 is twofold. First, we extend Study 1 results by testing the generalizability of mental model adjustment findings regarding the task domain, user expertise, and explanation presentation and examining whether the asymmetry we found for explanation-driven mental model adjustments in Study 1 is indeed a manifestation of the confirmation bias. Second, we explore if mental model adjustments spill over to related but disparate domains.

**3.2.1. Design.** Study 2 comprises four consecutive stages, where recruited real-estate experts estimated the listing price per square meter in Euro of apartments that we previously collected from a large online platform.<sup>16</sup> Participants saw ten apartment characteristics to make an informed guess and did not receive intermediate feedback. To reduce the task complexity and avoid informational overload, we fixed seven apartment characteristics across all stages, i.e., apartments only differed regarding the same three characteristics: Location (Frankfurt/Cologne), Balcony (Yes/No), Green voter share in the district (Below city average/City average/Above city average).<sup>17</sup> We provide screenshots of the interfaces from each stage in the supplementary material.

In Stage I, we elicited participants’ initial beliefs about the relationship between the three variable apartment characteristics and listing prices. Participants estimated the listing price of four random apartments with different combinations of the variable attributes by entering their marginal contributions to the price using a slider. Sliders ranged from minus to plus 2.500€ in steps of 50€. We initially set the marginal contributions and overall price estimation to 0€ and the average listing price (9600€), respectively. Participants additionally stated their confidence in the entered marginal contributions and the resulting price estimation on a five-point scale.

Stage II introduced our treatment variations. In all variations, participants estimated listing prices for eight random apartments with different combinations of variable attributes they did not

<sup>16</sup> We scraped data from a large online platform in February 2022. We collected observations for all apartments listed for sale in the seven major cities of Germany (“A-Cities”) and a medium-sized eastern German city (Chemnitz). We constructed a dataset consisting of eight apartment attributes and the listing price directly obtained from the platform, and two additionally collected features from public statistics. We provide summary statistics in the supplementary material (Table 6).

<sup>17</sup> Note: We selected these three characteristics for technical reasons regarding the ML model and based on the input from our industry partner. The notion is that these characteristics together are (i) sufficiently relevant to the prediction, and (ii) familiar/accessible to experts.

encounter in Stage I. In contrast to Stage I, participants directly entered the estimated listing price. As a reference point, they again observed the average listing price for an apartment. Participants stated their confidence on a five-point scale. In our baseline condition (NoAid), participants estimated the price without any aid. Participants in the AI condition observed opaque listing price predictions of a steady, i.e., non-learning, AI system trained on 4,975 collected observations.<sup>18</sup> In our XAI condition, in addition to observing these predictions, participants also saw numerically presented SHAP values for the three variable apartment characteristics, i.e., marginal contributions to the prediction in Euro. After they entered all eight listing price estimates, participants in treatments with decision support filled out a survey containing items on their trust, degree of reliance, and perceived transparency of the AI system (and explanations).

Stage III replicated Stage I to measure posterior beliefs. Independent of the condition, participants again made decisions without any aid for the same apartments.

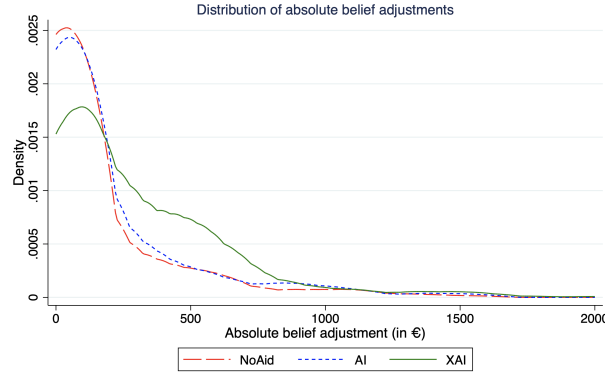
Finally, in Stage IV, participants estimated the listing price for one last apartment without any decision aid. Across participants, we varied the balcony and green voter attribute of the apartment, while the seven fixed attributes were identical to the previous listings. Most importantly, the apartment was in a midsize city in Eastern Germany (Chemnitz). For historical, demographic, and socioeconomic reasons, Chemnitz is very different from “A-cities” such as Frankfurt and Cologne, so the housing market is very different, too. Germans in general and real estate agents in particular are usually aware of this East-West disparity.<sup>19</sup> The study concluded with a questionnaire on participants’ socio-demographics.

**3.2.2. Results.** We report our results in three steps. First, we outline the experts’ belief adjustments from Stage I to Stage III. Second, we examine the occurrence of confirmation bias in these adjustment processes. Finally, we analyze experts’ listing price estimates in Stage IV.

**Mental model adjustments.** Figure 4 shows the distribution of absolute differences between experts’ beliefs about the marginal contribution of the three variable attributes before and after the treatment intervention. We show results for the NoAid, AI, and XAI conditions. The distributions for the NoAid and AI conditions are remarkably similar and skewed towards 0, indicating that experts frequently did not adjust beliefs. The distribution for XAI participants is considerably less right-skewed, i.e., they adjusted their beliefs across Stages I and III more. On average, NoAid, AI, and XAI participants adjusted their beliefs by 166.4€, 165.4€, and 299.1€, respectively. Only the differences between NoAid v. XAI ( $p < 0.01$ ,  $F$ -test), and AI v. XAI ( $p < 0.01$ ,  $F$ -test) conditions are statistically significant (see Table 24 in the supplementary material). Our notion is that real-estate

<sup>18</sup> The AI system is a Random Forest that achieves a performance of  $R^2 = 0.72$  on unseen test data. See the supplementary material for additional information.

<sup>19</sup> For instance, A-cities exhibit considerably higher average wages, more liberal political attitudes, and faster population growth (see, e.g., Cajias et al. 2020)



**Figure 4** Distribution of absolute belief changes.

Notes: We depict the distribution of experts' absolute belief adjustments across Stages I and III. We aggregate the belief adjustments over all apartment attributes. Different distributions show results separately for NoAid, AI, and XAI participants.

experts updated initially held mental models about the relationship between apartment attributes and listing prices as they encountered SHAP explanations. Contrasting our first study, we directly measure participants' prior and posterior beliefs about the contribution of distinct apartment characteristics to listing prices in Study 2. This design facet enables us to estimate mental model adjustments directly, leveraging the accepted framework by DeGroot (1974). Specifically, we assume that agent  $i$ 's posterior belief about the relationship of characteristic  $j$  and the listing price  $Post_{i,j} = a_{i,j} \cdot Prior_{i,j} + (1 - a_{i,j}) \cdot Expl_{i,j}$  is a weighted combination of the corresponding prior belief  $Prior_{i,j}$  and the personally observed explanation  $Expl_{i,j}$ .  $1 - a_{i,j}$  represents the extent of belief adaptation in the direction of the explanation, while  $a_{i,j}$  describes the anchoring of the previous belief. For instance, in the extreme case of  $1 - a_{i,j} = 1$ , individual  $i$  completely abandons her prior mental model and adopts the observed explanation as her new one. We estimate the weights  $(1 - a_i)$  and  $a_i$  for our three study conditions using a regression model comprising treatment interactions that has the following form:

$$\begin{aligned}
 Pos_{ijk} = & \beta_1 \cdot Pri_{ijk} + \beta_2 \cdot (AI_i \times Pri_{ijk}) + \beta_3 \cdot (Expl_i \times Pri_{ijk}) \\
 & + \beta_4 \cdot SV_{ij} + \beta_5 \cdot (AI_i \times SV_{ij}) + \beta_6 \cdot (Expl_i \times SV_{ij}) + \gamma_i + \delta_k + \epsilon
 \end{aligned} \tag{2}$$

$Pos_{ijk}$  and  $Pri_{ijk}$  respectively represent expert  $i$ 's posterior and prior beliefs about attribute  $j$ 's contribution to apartment  $k$ 's listing price in Euro. Most importantly,  $AI_i$  is a dummy variable indicating that expert  $i$  observed a prediction, while the dummy  $Expl_i$  equals 1 if a participant additionally observed explanations.  $SV_{ij}$  represents the average SHAP value for apartment attribute  $j$  of the apartments participant  $i$  encountered in Stage II.  $\gamma_i$  and  $\delta_k$  are expert and apartment controls.

On an individual level, model (2) estimates how observed SHAP values affected participants’ adjustments of beliefs about the relationship between a given characteristic and the listing price. It enables us to quantify the “stickiness” of prior beliefs ( $\beta_1 - \beta_3$ ) and “gravitational pull” of explanations ( $\beta_4 - \beta_6$ ), and directly test the occurrence of confirmation bias. Importantly, this estimation is only possible for Study 2, where we elicited prior and posterior beliefs about distinct feature-label relationships. In Study 1, we measured the ultimate investment decisions only and observed belief changes indirectly through changes in those decisions. As a result, we cannot individually quantify the impact of observed explanations on specific beliefs, nor can we analyze confirmation bias – a key contribution of our second study.

Dep. variable:	(1)	(2)
Posterior belief		
Prior belief ( $\beta_1$ )	0.634*** (0.060)	0.782*** (0.063)
Prior belief $\times$ AI ( $\beta_2$ )	0.070 (0.104)	-0.027 (0.084)
Prior belief $\times$ Expl. ( $\beta_3$ )	-0.276*** (0.084)	-0.240*** (0.075)
Avg. SHAP ( $\beta_4$ )	0.025 (0.040)	0.033 (0.039)
Avg. SHAP $\times$ AI ( $\beta_5$ )	0.078 (0.053)	0.083 (0.050)
Avg. SHAP $\times$ Expl. ( $\beta_6$ )	0.265*** (0.053)	0.249*** (0.052)
Fixed Effects	No	Yes
N	1836	1836
$R^2$	0.740	0.787

**Table 2** Posterior belief formation.

Notes: We depict results from OLS regression models with robust standard errors reported in parentheses. The dependent variable equals participants’ posterior belief about the marginal contribution of apartment attributes to the listing price in euros. The main independent variables of interest are participants’ prior beliefs, the average SHAP values for apartment attributes in Stage II, a dummy indicating that participants observed a prediction in Stage II (AI), a dummy indicating that participants observed explanations in Stage II (XAI), and interaction terms. We further control for the overall posterior listing price participants entered for the apartment and its interaction with treatment dummies, and the average prediction they observed in Stage II. In column (2), we additionally include individual and apartment fixed effects. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table 2 depicts regression results for model (2). Results show that in our NoAid and AI condition condition where participants did not observe explanations, SHAP values (unsurprisingly) have no significant explanatory power regarding posterior beliefs (see  $\beta_4$  and  $\beta_5$ ).<sup>20</sup> When participants did not obtain machine aid or only observed predictions, their prior and posterior beliefs were more than 60% positively correlated ( $\beta_1$  and  $\beta_2$ ), i.e., participants barely adjusted their beliefs. Only when participants observed explanations in addition to predictions did the displayed SHAP values have positive, statistically significant effects.  $\beta_6$  reveals that XAI participants significantly adjusted their beliefs in the direction of observed explanations. According to the estimate, posterior

<sup>20</sup> Note: the positive coefficient for  $\beta_5$  may be related to the fact that SHAP values and overall predictions are inextricably linked. Merely observing high (low) predictions may lead to adjustments of reported beliefs upward (downward), creating a positive, however, insignificant correlation with underlying SHAP values in the data.



beliefs resembled SHAP values more closely in the XAI treatment condition compared to the NoAid and AI conditions. Observing explanations also caused XAI participants’ posterior beliefs to resemble their prior significantly less ( $\beta_3$ ), i.e., prior beliefs became less “sticky” compared to the NoAid and AI conditions. In sum, these results suggest that observing SHAP explanations led participants to adjust their beliefs in the direction of explanations and abandon their priors. This insight corroborates our result 1.2 in Study 1 on an individual level, revealing that explanation-driven mental model adjustments also occur for experienced experts, who are arguably familiar with apartment traits and listing price predictions.<sup>21</sup>

**Confirmation bias.** In Study 1, we observed asymmetric mental model adjustments that are reminiscent of the confirmation bias. The design of Study 2 allows us to test for confirmation bias in mental model adjustment processes more directly by examining whether XAI participants’ adjustments depended on the alignment of explanations and prior beliefs.

We define that explanations confirmed an expert’s preconception about the price contribution of a specific apartment attribute if the prior and the observed average SHAP value for the corresponding attribute have the same sign. With this definition, observed explanations confirm prior beliefs in 49.6% of the cases.<sup>22</sup> We analyze differences in belief adjustments with respect to confirming and conflicting explanations using a modified version of model (2). Specifically, we are interested in whether the convergence of XAI participants’ posterior beliefs toward observed SHAP values only occurred when explanations confirmed prior beliefs. Therefore, we focus on the subsample of XAI participants allowing us to omit treatment dummies and interaction terms which facilitates the interpretation of results. Along the lines of model (2), we regress XAI participants’ posterior beliefs about the relationship between apartment characteristics and the listing price on their prior beliefs and observed SHAP values. Most importantly, we now add a dummy variable (*Confirm*) indicating whether explanations confirmed prior beliefs and its interaction with average SHAP values and prior beliefs as independent variables. The interaction  $Avg. SHAP \times Confirm$  will provide insights into whether the influence of observed SHAP values on belief adjustments depended on the alignment of explanations and prior beliefs – insights we can not obtain from study 1 using model (1).

Corroborating our interpretation of result 1.2 from Study 1, we find that explanation-driven belief adjustment processes depended on whether explanations confirmed or conflicted with prior beliefs. The estimate for the interaction term  $Avg. SHAP \times Confirm$  is positive and statistically

<sup>21</sup> Participants, on average, have worked in the real estate industry for 13.8 years and, on a scale from 1-10, report that their experience level in rating apartment listing prices equals 5.7.

<sup>22</sup> Our main insights are robust to defining more restrictively that explanations confirm priors if the absolute distance between the prior and the observed average SHAP value is smaller than the absolute distance between the prior and 0€ and, at the same time, smaller than the absolute distance between the prior and the closest extreme, i.e., +/- 2500€ (see Table 25 in the supplementary material).

Dep. variable:	(1)	(2)	(3)
Posterior belief	Overall	Low confidence beliefs	High confidence beliefs
Prior belief	0.492*** (0.091)	0.483*** (0.105)	0.496*** (0.136)
Avg. SHAP	0.303*** (0.043)	0.344*** (0.055)	0.145** (0.067)
Confirm	12.039 (27.949)	-10.838 (39.552)	115.724 (73.702)
Avg. SHAP $\times$ Confirm	0.166*** (0.059)	0.107 (0.077)	0.301*** (0.094)
N	708	481	222
$R^2$	0.746	0.725	0.843

**Table 3** Confirmation bias and posterior belief formation

Notes: We depict results from OLS regression models with individual and apartment fixed effects. We report robust standard errors reported in parentheses. The dependent variable equals XAI participants' posterior belief about the marginal contribution of apartment attributes to the listing price in euros. The main independent variables of interest are participants' prior beliefs, the average SHAP values for apartment attributes in Stage II, a dummy indicating that observed SHAP values in Stage II confirmed participants' priors – measured by an equal sign of prior beliefs and average SHAP values for a given attribute – and interaction terms. We further control for the overall posterior listing price participants entered for the apartment and the average prediction they observed in Stage II. Column (1) presents results for all decisions. Columns (2) and (3) respectively depict results for the shares of decisions where XAI participants report low and high confidence in their prior. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

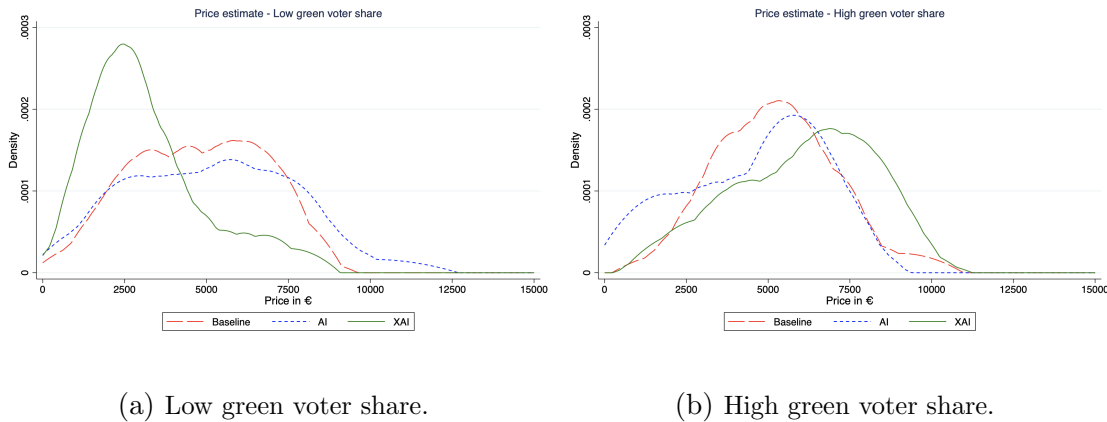
significant (see column (1) in Table 3). Following the estimate, posterior beliefs resembled observed SHAP values significantly more closely (about 50% more) if they confirmed their prior beliefs. Hence, consistent with confirmation bias, the belief adjustment was asymmetric regarding the confirmatory nature of explanations. If participants had updated beliefs rationally according to Bayes rule, the interaction term should be insignificant as Bayesian observers would not weight explanations conditional on their alignment with prior beliefs (Rabin and Schrag 1999).

To elaborate on the notion that these asymmetric belief adjustments are a manifestation of confirmation bias, we further consider the role of experts' confidence in their prior beliefs. Prior research shows that confirmation bias is strongest for entrenched beliefs (see, e.g., Pyszczynski and Greenberg 1987, Knobloch-Westerwick and Meng 2009). To test the existence of such heterogeneity, we consider experts' reported confidence in prior beliefs and define that an expert possessed low (high) confidence in a prior, if, on a 5-point scale, they reported a confidence level of less than 4 (at least 4). In columns (2) and (3) of Table 3, we respectively repeat the regression analysis reported in column (1) for the subsamples of low- and high-confidence prior beliefs.

Reported estimates provide further evidence that explanation-enabled mental model adjustments were subject to confirmation bias. According to the estimated coefficient of *Avg. SHAP*  $\times$  *Confirm*, for low-confidence priors, the influence of observed SHAP values on posterior beliefs did not depend on whether explanations confirmed prior beliefs (see column (2)). Considering the positive and significant estimate of *Avg. SHAP*, the belief updating was in line with Bayes rule. By contrast, for high-confidence priors, belief adjustments were highly sensitive to whether SHAP values confirmed priors (see column (3)). The estimate for *Avg. SHAP*  $\times$  *Confirm* suggests that the magnitude of the adjustment of high-confidence priors was about two times larger when observed explanations were in line with them.

**Result 2.1:** *Study 1 findings extend to expert users, SHAP explanations, and the domain of apartment price predictions: SHAP explanations led real-estate experts to adjust prior beliefs about the relation between apartment attributes and listing prices. Adjustment processes were subject to the confirmation bias.*

**Spillover effects.** While we observe that real-estate experts (asymmetrically) adjusted prior beliefs, all previously reported results pertain to the same market: participants observed SHAP explanations for the same two A-cities in Western Germany, for which we elicited prior and posterior beliefs. What remains open is whether explanation-driven belief adjustments spilled over to the listing price estimation for apartments in different markets. We put this idea to the test by examining the distribution of participants’ final price predictions for an apartment in a medium-sized eastern German city that is not an “A city”: Chemnitz.<sup>23</sup>



**Figure 5** Price distributions in Chemnitz.

Notes: We depict the distribution of experts’ listing price estimates in Chemnitz. Panel (a) and (b) depict price distribution for apartments in a district with a low and high share of green voters, respectively. Different distributions show results separately for NoAid, AI, and XAI participants.

Figure 5 shows the distribution of listing price estimates conditional on the share of green voters in the district for NoAid, AI, and XAI participants. The results indicate that observing explanations impacted participants’ price estimates for Chemnitz apartments in neighborhoods with high and low proportions of green voters. Panel (a) shows that the distribution of listing prices for an apartment in a district with a low green voter share is considerably more right-skewed for XAI than NoAid or AI participants, i.e., they estimate relatively low prices more frequently.

<sup>23</sup> Note that we did not include Chemnitz observations in the data to train the AI model. We conducted several analyses showing that the most important predictors for listing prices in Frankfurt and Cologne (cities in Stages I to III) differ considerably from listing price predictors in Chemnitz. Real-estate experts are arguably aware of the structural differences in apartment markets.

NoAid, AI, and XAI participants on average estimated a listing price of 4752€, 5141€, and 3140€, respectively. Only the differences between NoAid v. XAI and AI v. XAI are statistically significant in regression analyses ( $p < 0.05$ ,  $F$ -test, for both). The distribution of price estimates in districts with high shares of green voters has a stronger left-skew for XAI participants than their NoAid and AI counterparts (see Panel (b)). On average, NoAid, AI, and XAI participants estimated a listing price of 5231€, 4600€, and 6092€, respectively, for an apartment in a district with a high percentage of green voters. Again, we only find significant explanation effects ( $p < 0.1$ ,  $F$ -Test, for both). As one might expect, the direction of the difference in experts’ evaluation of the green voter share attribute is in line with explanations observed in Stage II: SHAP values indicated that in Frankfurt and Cologne, a high (low) share of green voters marginally contributes to listing prices by about +652€ (-613€). We do not find any effect for experts who only observed opaque predictions in Stage II.

To elaborate on these findings, we also perform a median split and analyze the subsamples of experts whose average absolute belief adjustment for the attribute “Green voter” is below and above the median. Consistent with the idea that belief spillover effects drive differences in listing price estimates in Chemnitz, experts who strongly adjusted their beliefs about the relevance of “Green voters” from Stage I to III drive our aggregate-level results. Note that we do not find significant treatment differences in the accuracy of participants’ listing price estimates, as measured by the absolute deviation from actual prices. Nevertheless, our results show that using XAI as a decision support tool in one market can affect aggregate listing prices in another market, which is not the case for opaque systems. This result demonstrates that XAI methods can link disparate decision-making tasks.

**Result 2.2:** *Pronounced explanation-driven belief adjustments spill over to experts’ listing price estimation in a fundamentally different market.*

In summary, our results from Study 2 (i) demonstrate the robustness of our results from Study 1 on mental model adjustments in terms of system user expertise, explanation representation, and decision domain, (ii) provide strong evidence that explanation-driven mental model adjustments are subject to confirmation bias, and (iii) that explanation-driven mental model adjustments generate significant spillover effects.

## 4. Discussion and Conclusion

We report results from two empirical studies that provide novel insights into the interplay between the employment of feature-based XAI methods and users’ cognitive processes. Our main contribution is the identification of considerable side effects of providing feature-based explanations – the

most popular form of XAI methods – on users’ situational information processing and mental models. We find that the latter effect (i) is subject to the confirmation bias so that misconceptions can persist and even accumulate, possibly leading to suboptimal decisions, and (ii) can create spillover effects into other decision domains. These overarching results suggest that the growing, partially legally required, employment of feature-based XAI methods opens a new channel through which AI systems may fundamentally reshape the way humans understand real-world relationships between features  $X$  and target variables  $Y$ . In the following, we discuss our results, present implications for organizations and society, and, based on the limitations of our studies, provide directions for future research.

**Discussion of results.** Study 1 demonstrates that the provision of explanations can situationally lead lay users to weigh features marked as important considerably more and put less emphasis on the overall prediction. Explanations also evoked asymmetric changes in lay users’ conceptions about the relationship between borrower traits and repayment inclinations that influence behaviors even when they do not observe explanations anymore, i.e., explanations affect mental models. Explanation-driven effects decreased lay users’ decision-making performance in our setting. Study 2 extended these results showing that even expert users in a considerably more applied domain adjusted mental models, that asymmetric mental model adjustments were a manifestation of the confirmation bias, and that mental model adjustments created spillover effects.

From a theoretical perspective, our results contribute to our understanding of the role of popular XAI methods in effective knowledge transfers from ML-based AI systems to human users. A key promise of modern AI systems is that the application of ML techniques will discover new knowledge from Big Data that has previously eluded even experienced experts (van den Broek et al. 2021, Berente et al. 2021). This “machine knowledge” is typically codified in the form of a complex predictive model that outperforms humans. We show that providing predictions alone is insufficient to achieve systematic knowledge transfers from AI systems to human users. In both our studies, neither laymen nor experts adapted their understanding of the relationships between features  $X$  and label  $Y$  according to “machine knowledge” when observing only opaque predictions. Merely in treatments where users also had access to explanations, they began to adapt their approach to solving the task so that it more closely matched the strategy of the AI system. Therefore, XAI methods appear to be a pivotal factor contributing to an effective channel through which AI systems can pass on their self-learned knowledge to human users. Crucially, feature-based XAI methods seem to induce an asymmetry in mental model adjustments: users adjust their beliefs more in the direction of observed explanations if they confirm rather than disconfirm their priors. This asymmetry contradicts with the updating behavior of a Bayesian observer who would neither over- nor underweight explanations conditional on them confirming or disconfirming prior beliefs. This

asymmetry occurred regardless of whether we provide graphically visualized LIME or numerically represented SHAP explanations. It, therefore, seems as if additive feature-based explanations more generally evoke cognitive processes leading users to learn from the machine selectively. Researchers across disciplines commonly refer to such an asymmetry as confirmation bias (Yin et al. 2016). Study 2 provides consistent evidence that explanation-driven knowledge transfers from an AI to a human similarly suffer from confirmation bias as knowledge transfers in the human-to-human domain. For example, confidence in prior conceptions and their difference from the new information moderate confirmation bias (Pyszczynski and Greenberg 1987). Similar to learning from other humans, users seem unwilling to internalize potentially helpful, XAI-channeled machine knowledge if it is inconsistent with what they already, perhaps incorrectly, believe to be true. From the perspective of the Mental Models framework, individuals more frequently engage in maintaining rather than in building mental models of the relationships between features and labels. One reason for this effect could be the need to attain or maintain a high level of self-esteem (Klayman 1995), leading users to focus inappropriately on explanations that make them feel competent. In other words, they may derive a positive intrinsic benefit from being in the right (e.g., Gilad et al. 1987). From this perspective, people may misuse the XAI as a tool to enhance their self-esteem. If left unaddressed, the asymmetric adaptation of mental models by humans may prevent modern (X)AI applications from fulfilling their promise of making humans smarter, which (ironically) may also hinder the further development of AI applications by humans.

Interpreting our results in the light of the model by Agrawal et al. (2019) yields another theoretical insight regarding the ramifications of XAI. Our results indicate that users' willingness to follow XAI predictions depends on whether the explanations conform with their mental models. One way to rationalize this behavior is that their objective function includes a component that accounts for experiencing some positive (negative) intrinsic utility when obtaining a signal that their mental model may (not) be accurate (see, e.g., Gilad et al. 1987, Festinger 1962, Harmon-Jones 2019). In the model by Agrawal et al. (2019), AI systems make predictions about uncertain states of the world that relate to the profitability of taking specific actions. Human users, in turn, assess the expected payoffs associated with specific actions, i.e., make judgments. Our results suggest that human judgment in this model encompasses not only the material consequences of an action, but also the psychological impact of receiving a signal that implicitly shows whether current mental models are correct. If explanations reveal that the AI system arrived at a prediction in a way that contradicts their held mental models, taking an action that follows this prediction effectively constitutes a signal to oneself that the current mental model is incorrect, creating psychological distress, e.g., in the form of a cognitive dissonance (Harmon-Jones 2019). This mental toll may lead

users not to follow the prediction in the first place. Conversely, users may follow unreliable predictions more often if the explanations are consistent with their current mental models because doing so provides a psychologically valuable self-signal that they are in the right (see, e.g., Gilad et al. 1987). Against this background, users’ inclination to follow predictions of an XAI system, and thus their ultimate decisions and gains, is subject to greater variance than with a black-box AI. That is because users’ propensity to follow predictions depends on the consistency of the explanations with their mental models.

Another theoretical contribution of our work is to show the potential of feature-based XAI to link different decision domains by influencing users’ beliefs about the feature-label relationship. Study 2 results show that observing explanations for listing price predictions for apartments in Market A influenced the price estimation of experts in a different Market B, where the learned pattern does not exist and they did not have access to XAI decision support. We find that listing prices estimated by experts who observed explanations differed significantly from those estimated by experts who either had no decision aid or only observed opaque predictions. This spillover effect seems to occur due to the adjustment of mental models that experts draw upon in both situations. Therefore, as an unintended side effect, increasing public and private efforts to promote the use of XAI methods may extend the already significant influence of AI systems from areas where we interact with them (Rahwan et al. 2019) to areas where such systems are not in use. Feature-based XAI methods’ potential to link different domains is particularly concerning given recent evidence on their susceptibility to intentional manipulation and adversarial attacks (Lipton 2018). Many modern XAI methods, including LIME and SHAP, optimize fidelity, i.e., ensure that explanations accurately mimic the predictions of the black box model. However, even small perturbations of the input data (e.g., deliberate manipulation and measurement errors) can lead to considerably different explanations for identical predictions, i.e., depict different feature-label relations (Ghorbani et al. 2019, Lakkaraju and Bastani 2020). The potential instability of explanations allows manipulating user behaviors. Following our results, the creation of misleading explanations may not only affect users’ trust in the AI system (Lakkaraju and Bastani 2020), but also lead to an (asymmetric) adjustment of mental models that affect users’ decision making beyond the XAI augmented decision at hand. Specifically, the depiction of certain feature-label relationships that are not present can evoke inappropriate mental model adjustments that, given the documented asymmetry, will cause users who already believe these patterns to be true, to feel vindicated and reinforce these beliefs. In general, the documented spillover effects may magnify the reach and impact of intentional manipulations of explanations, increasing deceiving parties’ incentive to do so.

**Implications.** Reported results have important practical implications for organizations and policy-makers. Our finding that XAI can change human thinking points to potential pitfalls for companies

that want, or have to, use XAI. Consider a company that plans to implement XAI methods to explain to its employees why an AI system makes certain predictions. As Study 1 shows, providing explanations in addition to predictions may draw users’ attention excessively to the explanations, to the detriment of the prediction itself. Users may place too much emphasis on individual explanations that confirm their prior beliefs, rather than adhering to the overall prediction. As a result, employees’ decision-making performance for the task at hand may deteriorate, which is in line with evidence from related research (see, e.g., Poursabzi-Sangdeh et al. 2021). In domains where explanations are becoming a regulatory standard, managers need to take such potential downsides into account and contemplate the ramifications of implementing explainability measures. Following our results, managers who, in the future, are obliged to put XAI methods in place, should not take these steps too lightly. From a business perspective, our documented downsides of explainability could render the continued use of AI-based decision support systems unattractive. Considering that AI systems are often deeply interwoven with business processes, this XAI-driven discontinuance may entail considerable organizational change. As a result, managers may be well-advised to assess potential inconsistencies between the AI system’s internal logic and employees’ understanding of the task it supports before rolling out explainability measures. This puts managers in a position to evaluate the magnitude of the potential downside of explainability and employ countermeasures. For example, managers may obviate confirmation bias by openly discussing explanations that conflict with employees’ mental models and showcasing arguments in support of the explanation.

Another pitfall for companies concerns the transfer of knowledge from AI systems to human users. As Study 2 shows, even experts can overgeneralize learned feature-label relationships that are only applicable in the context in which they interact with the system. With the confirmatory learning from explanations, existing differences in employees’ initial conceptions may lead to differences in how they collaborate with and what they learn from the XAI, e.g., fostering the biased weighting of certain information. From this perspective, providing explanations might decrease individual level noise in the decision-making process (Kahneman et al. 2021), because individuals’ decisions become more consistent. This is in line with Fügner et al. (2021b) who find decisions to be increasingly consistent among users engaging with opaque predictions. On a more aggregate level, however, our results suggest that explained predictions may additionally foster differences in the decision-making process across subgroups of users that possess heterogeneous priors. As a consequence, the variation of decisions on a group-level can grow. As pointed out by Kahneman et al. (2021), variation in decisions can substantially contribute to errors and ultimately harm business performance. Consider our previous example of loan officers. XAI may cause loan approval decisions to increasingly depend on the particular employee – with idiosyncratic mental models – assessing the applicant’s creditworthiness. This increase in loan approval variation may create considerable



business, legal, and reputational risks. Against this background, managers should closely monitor the introduction of XAI to identify a possible increase in decision variance. For instance, managers could complement XAI with “noise audits” and the development of “reasoned rules” (as proposed by Kahneman et al. 2021) to overcome the hidden costs of XAI-driven increases in inconsistent decision-making.

From a societal perspective, our results indicate that broad, indiscriminate implementation of XAI methods may create unintended downstream ramifications. Our finding that XAI can lead users to adjust mental models in a confirmatory way and carry over learned patterns to other domains may, in an extreme case, may foster discrimination and social divisions. Assume all recruiters start to collaborate with an XAI system to support hiring decisions. For example, a subgroup of recruiters may discriminate against women because they believe female applicants to be less productive on the job. If the XAI (occasionally) provides local explanations that depict being female as negative evidence for high future performance, the subgroup that statistically discriminates based on gender will readily reinforce its prior belief, i.e., engage in mental model maintenance. As a result, these recruiters may become more biased and less noisy in their behavior as they hire female applicants consistently less. Given the spillover effects we find, they may even carry over their strengthened conceptions about women’s productivity to other jobs, further reinforcing discriminatory patterns. Additionally, because non-discriminating recruiters will most likely refrain from adjusting their mental model, i.e., not engage in mental model building, social divisions among recruiters may develop and accumulate along the lines of gender biases. Hence, without any malicious intent, the broad employment of XAI may ironically foster human discriminatory tendencies and divide social groups. Notably, with the possibility to manipulate explanations, deceiving third parties could also intentionally cause explanations to exhibit specific prediction contributions for sensitive attributes such as race, gender, or age. This effect could lead human users who already hold prejudices, stereotypes, or discriminatory tendencies to reinforce their views, which could promote certain political agendas, for example.

**Limitations and future research.** As with any other research study, ours is not without limitations. In light of increasing regulatory requirements and private initiatives, we believe that these limitations open up fruitful avenues for future research. One limitation of our work concerns the lack of feedback on the decision outcomes and, thus, the performance of the AI system. In both our studies, we did not provide feedback for two reasons. First, it adds a considerable layer of complexity that impedes the measurement and interpretation of isolated explanation-driven effects on users’ cognitive processes. Second, in practice, many AI-supported decisions do not yield immediate feedback, or only yield feedback for some of the predictions. Hence, users have to interact with the

system without learning its prediction accuracy, at least for a certain period. Examples include hiring decisions supported by an on-the-job performance predicting AI system, investment decisions supported by a return predicting AI system, and drug treatment decisions supported by an effectiveness predicting AI system. Consequently, explanations may alter users' situational information processing and mental models before feedback on system performance arrives. Nonetheless, we strongly encourage future research to examine the role of feedback as it may introduce unexpected dynamics in the cognitive effects we document. For instance, the (selective) reinforcement of their mental models through explanations, may lead users to be more forgiving and maintain trust in the AI system, even if they eventually see it making mistakes. In this way, the interaction between feedback and explanations might constitute a factor contributing to unwarranted algorithm appreciation (Logg et al. 2019), leading users to rely on incorrect outcomes blindly. Additionally, people's adjustments of the situational information processing and existing mental models possibly depend on the extent to which the XAI system's predictions outperform their own. If users learn that an XAI system's predictions perform considerably better than their subjective ones, the magnitude of reported confirmation biases may vary. Conversely, when users' predictions are better than the XAI, their confirmation bias might be even stronger. Future research could examine to what extent our reported effects, at the intensive margin, depend on users' perceptions about differences in their own and the XAI system's predictive performance.

Another limitation of our work originates from letting participants interact with local, feature-based XAI methods. We opted to employ these explanations because they are already widely in use in practice and because there are arguments that feature-based explanations on an individual level are necessary to comply with (upcoming) regulatory requirements (Goodman and Flaxman 2017). Yet, there exist other forms of explanations, e.g., global feature-based explanations or even example-based explanations. Even though an investigation and comparison of the interplay between different forms of explanations and cognitive processes are beyond the scope of this paper, it is worthwhile for future research to explore whether, and if so why, the effects we document would change if users (additionally) obtain other forms of explanations. Consider, for instance, global explanations. While local explanations help understand why an AI system produces a prediction on a case-by-case basis, global explanations reveal important high-level patterns and non-linearities in the system's logic. Such global explanations effectively aggregate individual-level information for the user and help to understand the system's overall logic. By taking over this information aggregation task, global explainability could mitigate concerns about the selective processing of isolated local explanations that arguably contribute to the occurrence of confirmation bias. Additionally, the global representation may facilitate comparison and reflection processes which ultimately improves the transfer of knowledge from the AI system to the user.

**Conclusion.** A concluding remark is worth making. Of course, our work is not meant to be an argument, let alone a plea, against making “black box” AI systems more explainable or transparent. Instead, we comprehend our findings as a warning that the indiscriminate use of modern XAI methods as an isolated measure may lead to unintended, unforeseen problems because it creates a new channel through which AI systems can affect human behaviors across domains. The pervasive human inclination to process information in a way that confirms their preconceptions while ignoring potentially helpful yet conflicting information needs addressing if explainability is to become an effective means to combat accountability, transparency, and fairness issues without creating adverse second-order effects. For instance, one might restrict the provision of explanations of sensitive features for end-users of the system and only use them to ensure the proper and unbiased functioning of the AI system during the development process. Additionally, it might be important to provide developers and data scientists with cognitive awareness trainings to make them more sensitive to their own biased mental processes.

## References

- Abdel-Karim BM, Pfeuffer N, Carl V, Hinz O (2022) How AI-based systems can induce reflections: The case of ai-augmented diagnostic work. *MIS Quarterly* conditionally accepted.
- Abdel-Karim BM, Pfeuffer N, Rohde G, Hinz O (2020) How and what can humans learn from being in the loop? *German Journal of Artificial Intelligence* 34(2):199–207.
- Agarwal R, Dhar V (2014) Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research* 25(3):443–448.
- Agrawal A, Gans JS, Goldfarb A (2019) Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy* 47:1–6.
- Ai C, Norton EC (2003) Interaction terms in logit and probit models. *Economics letters* 80(1):123–129.
- Alavi M, Marakas GM, Yoo Y (2002) A comparative study of distributed learning environments on learning outcomes. *Information Systems Research* 13(4):404–415.
- Ban GY, El Karoui N, Lim AE (2018) Machine learning and portfolio optimization. *Management Science* 64(3):1136–1154.
- Bauer K, Hinz O, van der Aalst W, Weinhardt C (2021) Expl(AI)n it to me—explainable AI and information systems research. *Business & Information Systems Engineering* 63(2):79–82.
- Berente N, Gu B, Recker J, Santhanam R (2021) Managing artificial intelligence. *MIS Quarterly* 45(3):1433–1450.
- Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games and Economic Behavior* 10(1):122–142.
- Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, Ghosh J, Puri R, Moura JM, Eckersley P (2020) Explainable machine learning in deployment. *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Brewer WF (1987) Schemas versus mental models in human memory. *Modelling cognition* 187–197.
- Bussone A, Stumpf S, O’Sullivan D (2015) The role of explanations on trust and reliance in clinical decision support systems. *International Conference on Healthcare Informatics*.
- Cabral TS (2021) AI and the right to explanation: Three legal bases under the GDPR. *Data Protection and Privacy, Volume 13: Data Protection and Artificial Intelligence* 13:29–56.
- Cajias M, Freudenreich P, Freudenreich A, Schäfers W (2020) Liquidity and prices: A cluster analysis of the german residential real estate market. *Journal of Business Economics* 90(7):1021–1056.
- Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* 19(1):7–42.
- Case N (2018) How To Become A Centaur. *Journal of Design and Science*  
<https://jods.mitpress.mit.edu/pub/issue3-case>.

- Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809–825.
- Chatterjee S, Sarker S, Valacich JS (2015) The behavioral roots of information systems security: Exploring key factors related to unethical it use. *Journal of Management Information Systems* 31(4):49–87.
- DeGroot MH (1974) Reaching a consensus. *Journal of the American Statistical Association* 69(345):118–121.
- Dellermann D, Ebel P, Söllner M, Leimeister JM (2019) Hybrid intelligence. *Business & Information Systems Engineering* 61(5):637–643.
- Dhaliwal JS, Benbasat I (1996) The use and effects of knowledge-based system explanations: Theoretical foundations and a framework for empirical evaluation. *Information Systems Research* 7(3):342–362.
- Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114–126.
- Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.
- Dijkstra JJ (1999) User agreement with incorrect expert system advice. *Behaviour & Information Technology* 18(6):399–411.
- Dodge J, Liao QV, Zhang Y, Bellamy RK, Dugan C (2019) Explaining models: an empirical study of how explanations impact fairness judgment. *International conference on Intelligent User Interfaces*.
- Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*.
- Erlei A, Nekdem F, Meub L, Anand A, Gadiraju U (2020) Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. *AAAI Conference on Human Computation and Crowdsourcing*.
- EU (2016) 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.
- EU (2021) 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.
- Fehr E, Fischbacher U (2003) The nature of human altruism. *Nature* 425(6960):785–791.
- Festinger L (1962) Cognitive dissonance. *Scientific American* 207(4):93–106.
- Fügener A, Grahl J, Gupta A, Ketter W (2021a) Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research* forthcoming.
- Fügener A, Grahl J, Gupta A, Ketter W (2021b) Will humans-in-the-loop become borgs? merits and pitfalls of working with AI. *MIS Quarterly* 45(3b):1527–1556.

- Garreau D, Luxburg U (2020) Explaining the explainer: A first theoretical analysis of lime. *International Conference on Artificial Intelligence and Statistics*.
- Ge R, Zheng Z, Tian X, Liao L (2021) Human–robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Information Systems Research* 32(3):774–785.
- Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. *AAAI Conference on Artificial Intelligence*.
- Gilad B, Kaish S, Loeb PD (1987) Cognitive dissonance and utility maximization: A general framework. *Journal of Economic Behavior & Organization* 8(1):61–73.
- Goldstein IM, Lawrence J, Miner AS (2017) Human-machine collaboration in cancer and beyond: The centaur care model. *JAMA Oncology* 3(10):1303–1304.
- Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine* 38(3):50–57.
- GoogleAI (2019) Responsible AI practices - interpretability. <https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability>, accessed: 2022-03-08.
- Gramegna A, Giudici P (2021) SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence* 4:752558.
- Gregor S (2006) The nature of theory in information systems. *MIS Quarterly* 30(3):611–642.
- Gregor S, Benbasat I (1999) Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly* 23(4):497–530.
- Gunning D, Stefk M, Choi J, Miller T, Stumpf S, Yang GZ (2019) XAI—explainable artificial intelligence. *Science Robotics* 4(37):eaay7120.
- Harmon-Jones EE (2019) *Cognitive dissonance: Reexamining a pivotal theory in psychology* (American Psychological Association).
- Hemmer P, Schemmer M, Vössing M, Kühl N (2021) Human-AI complementarity in hybrid intelligence systems: A structured literature review. *Pacific Asia Conference on Information Systems (PACIS)*.
- Hoffman M, Kahn LB, Li D (2018) Discretion in hiring. *The Quarterly Journal of Economics* 133(2):765–800.
- Holt CA, Smith AM (2009) An update on bayesian updating. *Journal of Economic Behavior & Organization* 69(2):125–134.
- Ji-Ye Mao IB (2000) The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems* 17(2):153–179.
- Johnson-Laird PN, Goodwin GP, Khemlani SS (2017) Mental models and reasoning. *The Routledge international handbook of thinking and reasoning*, 346–365 (Routledge).
- Jones NA, Ross H, Lynam T, Perez P, Leitch A (2011) Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society* 16(1).

- Jussupow E, Benbasat I, Heinzl A (2020) Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *European Conference on Information Systems (ECIS)*.
- Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians’ decision-making process with artificial intelligence. *Information Systems Research* 32(3):713–735.
- Kahneman D, Sibony O, Sunstein CR (2021) *Noise: A flaw in human judgment* (Little, Brown).
- Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J (2020) Interpreting interpretability: understanding data scientists’ use of interpretability tools for machine learning. *CHI Conference on Human Factors in Computing Systems*.
- Klayman J (1995) Varieties of confirmation bias. *Psychology of learning and motivation* 32:385–418.
- Kleinmuntz B (1990) Why we still use our heads instead of formulas: toward an integrative approach. *Psychological bulletin* 107(3):296.
- Knobloch-Westerwick S, Meng J (2009) Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research* 36(3):426–448.
- Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. *International Conference on Machine Learning (ICML)*.
- Lakkaraju H, Bastani O (2020) “How do i fool you?” Manipulating user trust via misleading black box explanations. *AAAI/ACM Conference on AI, Ethics, and Society*.
- Lakkaraju H, Kamar E, Caruana R, Leskovec J (2019) Faithful and customizable explanations of black box models. *AAAI/ACM Conference on AI, Ethics, and Society*.
- Lim KH, Ward LM, Benbasat I (1997) An empirical study of computer system learning: Comparison of co-discovery and self-discovery methods. *Information Systems Research* 8(3):254–272.
- Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3):31–57.
- Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.
- Lu Z, Yin M (2021) Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. *CHI Conference on Human Factors in Computing Systems*.
- Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Conference on Neural Information Processing Systems (NIPS)* .
- Malle BF (2006) *How the mind explains behavior: Folk explanations, meaning, and social interaction* (MIT press).
- Meske C, Bunde E, Schneider J, Gersch M (2022) Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management* 39(1):53–63.

- MetaAI (2021) Facebook’s five pillars of responsible AI. <https://ai.facebook.com/blog/facebook-five-pillars-of-responsible-ai/>, accessed: 2022-03-08.
- Miettinen T, Kosfeld M, Fehr E, Weibull J (2020) Revealed preferences in a sequential prisoners’ dilemma: A horse-race between six utility functions. *Journal of Economic Behavior & Organization* 173:1–25.
- Molnar C (2020) *Interpretable machine learning: A Guide for Making Black Box Models Explainable*.
- Nori H, Jenkins S, Koch P, Caruana R (2019) InterpretML: A unified framework for machine learning interpretability. *arXiv:1909.09223*.
- Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H (2021) Manipulating and measuring model interpretability. *CHI Conference on Human Factors in Computing Systems*.
- Pyszczynski T, Greenberg J (1987) Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. *Advances in Experimental Social Psychology*, volume 20, 297–340.
- Rabin M, Schrag JL (1999) First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics* 114(1):37–82.
- Rader E, Cotter K, Cho J (2018) Explanations as mechanisms for supporting algorithmic transparency. *CHI Conference on Human Factors in Computing Systems*.
- Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO, et al. (2019) Machine behaviour. *Nature* 568(7753):477–486.
- Ribeiro MT, Singh S, Guestrin C (2016) “Why should I trust you?” Explaining the predictions of any classifier. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Rico-Juan JR, de La Paz PT (2021) Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in alicante, spain. *Expert Systems with Applications* 171:114590.
- Rosenfeld A, Richardson A (2019) Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems* 33(6):673–705.
- Rouse WB, Morris NM (1986) On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin* 100(3):349.
- Schanke S, Burtch G, Ray G (2021) Estimating the impact of “humanizing” customer service chatbots. *Information Systems Research* 32(3):736–751.
- Schön DA (2017) *The reflective practitioner: How professionals think in action* (Routledge).
- Senoner J, Netland T, Feuerriegel S (2021) Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science* forthcoming.
- Shapley LS (1953) A value for n-person games. *Contributions to the Theory of Games (AM-28), Volume II* (Princeton University Press).



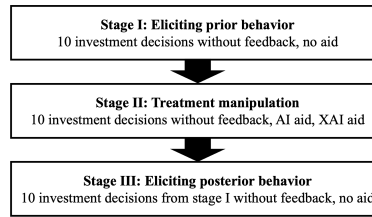
- Teodorescu MH, Morse L, Awwad Y, Kane GC (2021) Failures of fairness in automation require a deeper understanding of human-ml augmentation. *MIS Quarterly* 45(3b):1483–1499.
- Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, Janda M, Lallas A, Longo C, Malvehy J, et al. (2020) Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26(8):1229–1234.
- van den Broek E, Sergeeva A, Huysman M (2021) When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly* 45(3):1557–1580.
- Vandenbosch B, Higgins C (1996) Information acquisition and mental models: An investigation into the relationship between behaviour and learning. *Information Systems Research* 7(2):198–214.
- Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76:89–106.
- Wang W, Benbasat I (2007) Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23(4):217–246.
- Willison R, Warkentin M (2013) Beyond deterrence: An expanded view of employee computer abuse. *MIS Quarterly* 37(1):1–20.
- Yang F, Huang Z, Scholtz J, Arendt DL (2020) How do visual explanations foster end users’ appropriate trust in machine learning? *International Conference on Intelligent User Interfaces*.
- Yin D, Mitra S, Zhang H (2016) Research note—when do consumers value positive vs. negative reviews? an empirical investigation of confirmation bias in online word of mouth. *Information Systems Research* 27(1):131–144.

## Supplementary material

Our Supplementary material comprises four parts. In parts 1 and 2 (**Study 1: Study design** and **Study 2: Study design**), we provide detailed descriptions on the design of Study 1 and 2, respectively. In part 3 (**Study 1: Analyses**), we report the analyses for Study 1 results that we refer to in the main text. In part 4 (**Study 2: Analyses**), we do so for the Study 2 results.

### Study 1: Experimental design

**Overview.** The experiment comprises 3 consecutive stages (see Figure 6 for an overview). In each stage, participants repeatedly engaged in a modified version of the one-shot investment game (Berg et al. 1995) that is detailed in the following.

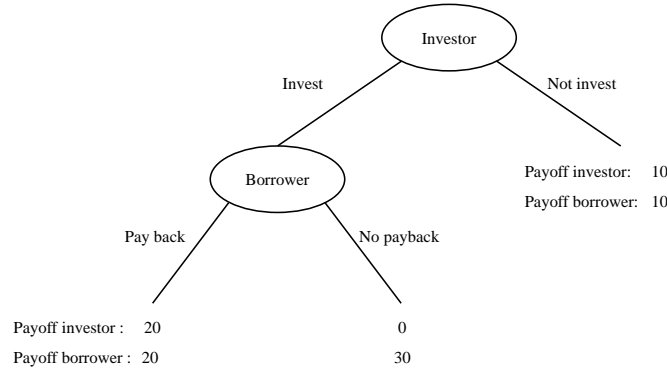


**Figure 6** Sequence of the experiment

Notes: Sequence and overview of the 3 different stages in the experiment.

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 7).

This investment game mimics the fundamental structure of many sequential, strategic decisions under uncertainty (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 paid 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 7** Investment game structure

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

In a nutshell, our experiment works as follows. There are three subsequent stages, with every single stage being individually incentivized. In Stage I we elicit participants’ prior investment behavior by letting them make several investment decisions without intermediary feedback. In Stage II, 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. Stage III mirrors Stage I, allowing us to elicit investors’ posterior behavior. We show the developed interfaces in Figure 8 and Figure 9, respectively. 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.

In addition to the three stages detailed above, we additionally measure participants’ prior and posterior preferences to observe borrower characteristics. We measure the preferences right before and after Stage II and use them for robustness and consistency checks.

**Details on borrowers, the AI, and 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). 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.

**Part 1 - Round 1 of 10**

Current person's personal characteristics	
The other person's biological sex:	Male
Whether the other person has younger siblings:	No
Whether the other person has older siblings:	No
Level of the other person's patience:	High
Level of the other person's approachability:	High
Propensity of the other person to be warm and considerate towards others:	Very high
Level of the other person's competitiveness:	Medium
Propensity of the other person to become upset/ stressed:	Low
Level of the other person's openness to new experiences:	Very low
Level of the other person's conscientiousness:	High

**Please make your decision and click on the "Next"-Button (appears after 5 seconds)**

**What do you want to do:**

- ☐ **Keep your 10 monetary units**  
☐ **Transfer your 10 monetary units**

Next

**Figure 8 Interface of Study 1 in Stage I and III.**

Notes: We show the interface developed to let participants in Study 1 make investment decisions in Stage I and III, i. e., without any aid.

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 ( $n=50$ ).<sup>24</sup> 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 supplementary material). The randomly drawn 50 observations serve as the population of borrowers with whom participants play in our study. We choose to select 50 borrowers, even though each participant only interacts with 32 borrowers (the same 10 in Stages I and III, 20 in Stage II, and 2 for eliciting prior

<sup>24</sup> Note: a Kolmogorov-Smirnov test cannot reject the hypothesis that both sets stem from the same underlying population  $p = 0.781$

**Part 3 - Round 1 of 20**

**Prediction by Machine Learning System**  
**about the other person's propensity to reciprocate a transfer**

**You will most likely receive 0 monetary units**, if you initially make a transfer  
 (i.e. the other person will most likely NOT reciprocate your transfer).

	Current person's personal characteristics	Importance of characteristic for current prediction
The other person's biological sex:	Female	<div style="width: 20px; height: 10px; background-color: green;"></div>
Whether the other person has younger siblings:	Yes	<div style="width: 5px; height: 10px; background-color: black;"></div>
Whether the other person has older siblings:	Yes	<div style="width: 5px; height: 10px; background-color: red;"></div>
Level of the other person's patience:	Low	<div style="width: 20px; height: 10px; background-color: green;"></div>
Level of the other person's approachability:	Medium	<div style="width: 10px; height: 10px; background-color: red;"></div>
Propensity of the other person to be warm and considerate towards others:	Low	<div style="width: 5px; height: 10px; background-color: black;"></div>
Level of the other person's competitiveness:	Very high	<div style="width: 30px; height: 10px; background-color: red;"></div>
Propensity of the other person to become upset/ stressed:	Very high	<div style="width: 20px; height: 10px; background-color: green;"></div>
Level of the other person's openness to new experiences:	High	<div style="width: 10px; height: 10px; background-color: red;"></div>
Level of the other person's conscientiousness:	Very high	<div style="width: 5px; height: 10px; background-color: black;"></div>

**Please make your decision and click on the "Next"-Button (appears after 5 seconds)**

**What do you want to do:**

- ☐ **Keep your 10 monetary units**  
☐ **Transfer your 10 monetary units**

Next

**Figure 9**      **Interface of Study 1 in Stage II, XAI treatment**

Notes: We show the interface developed to let participants in the XAI treatment in Study 1 make investment decisions in Stage II. Notably, in the AI treatment, participants did not observe the graphically visualized explanations.

and posterior preferences). Our intention is to create variation on the side of the borrower so that participants not always interact with the same 32 borrowers which might bias our results.

Investors in our online study always observe these ten borrower characteristics before making their decision (see Table 4 for an overview).

		Distribution of continuous values				
Item		Very low	Low	Medium	High	Very High
1.	Big 5: Openness	6%	18%	26%	26%	24%
2.	Big 5: Conscientiousness	-	4%	16%	48%	32%
3.	Big 5: Extraversion	2%	14%	24%	20%	40%
4.	Big 5: Agreeableness	-	4%	8%	34%	54%
5.	Big 5: Neuroticism	16%	32%	16%	26%	10%
6.	Competitiveness	12%	14%	14%	22%	38%
7.	Patience	6%	28%	16%	26%	24%

		Distribution of binary values	
Item		No	Yes
8.	Gender (male)	40%	60%
9.	Person has younger siblings	50%	50%
10.	Person has older siblings	46%	54%

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

Notes: We show the features used to train the ML model together with the distribution of values for the sample of observations used in the experiment.

The main motivation for choosing these borrower characteristics in Study 1 is that we wanted to develop a high-performing AI model that uses relevant input features, over which participants do not hold strong beliefs that they bring into the controlled environment of the experiment. From a theoretical point of view, extensive literature in the field of Economics and Psychology documents the strong relationship between the used personality traits and social preferences, including positive reciprocity that plays a pivotal role in the motivation of second movers to make a repayment in investment/trust games (see, e.g., Dohmen et al. 2009, Becker et al. 2012).<sup>25</sup>

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. 2016). LIME is one of the most popular and widely used explainability techniques as of today (see e.g., Gramegna and Giudici 2021, 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

<sup>25</sup> Using the standard ten-fold cross-validation, the model achieves an average performance of about 74% accuracy.

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 %.

Following the standard approach suggested by Ribeiro et al. (2016), 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.

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.

**Stage I.** In Stage I, participants played 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.

**Stage II.** Stage II 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.<sup>26</sup>

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. 2016). In our treatment condition, participants observe LIME explanations for

<sup>26</sup> 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.



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 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.

**Stage III.** Finally, in Stage III, 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. Notably, we do not explicitly explain this detail to participants in order to avoid anchoring their choice.

**Preference measures.** In addition to the three main stages, we additionally measure participants’ prior and posterior preferences to observe borrower characteristics.

We measure the prior preferences right before Stage II. Participants play one investment game against a random borrower from the player set whom they do not encounter in the main stages. In contrast to the main stages, 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.

We measure the posterior preferences right after Stage II. Participants again play one investment game that mirrors the one from eliciting the prior preferences, but against a different random

borrower. Again, the investment decision is payoff relevant in any case and participants do not receive feedback.

The preference measures are intended as a robustness check to test whether the presence of explanations affected participants’ initially most pronounced preferences. Specifically, letting participants choose three characteristics allows us to obtain ranking preferences by observing specific borrower traits, in an arguably credible and incentive-compatible way. We restricted the choice to three features because we wanted to (i) have a relatively small choice set that motivated participants to contemplate seriously which trait they preferred to observe, and (ii) decrease the likelihood that participants could choose larger combinations of features making sense only together, i.e., reducing concerns about the complementarity of isolated preferences over traits.

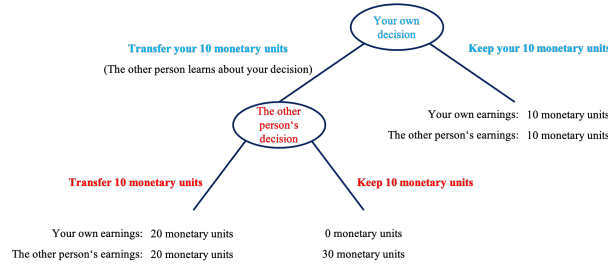
**Completion** After participants have made all investment decisions, the experiment ends with a questionnaire containing items on participants’ socio-demographics and social preferences. Participants’ answers serve as controls for some of our regression analyses. At the end of the experiment, we inform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

**Experimental summary.** Overall, 607 individuals participated in our study (301 Treatment condition and 306 Baseline condition). 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. We match participants’ investor decisions with corresponding borrower decisions to determine payoffs according to the previously outlined structure. For each of the three main stages where participants make multiple investment decisions, we randomly select one of the rounds. We informed participants of this randomness in every single stage. 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. 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. We inform participants about this payoff procedure in the instructions. Under the reasonable assumption that participants maximize expected utility, these (probabilistic) payouts to borrowers ensure incentive compatibility regarding preferences over borrower’s material well-being that investor decisions affect.

**Instructions.** In the following, we present the instructions of Study 1. Please note that Stage I, II, and III correspond to Parts 1, 3, and 5. The elicitation of prior and posterior preferences correspond to Parts 2 and 4, respectively.

### Part 1

In part 1 of the experiment, you play 10 rounds of a game that has the following structure (see the figure for an illustration).



At the beginning of every round, you are randomly matched with a new anonymous person from another study. Both you and the other person receive 10 monetary units. Your task is always the same: You start making a decision about whether you want to keep your 10 monetary units or transfer all of them to the other person. Note: You can not transfer only a part of your endowment.

Keeping and transferring your monetary units has the following consequences:

**Keeping your 10 monetary units:** If you decide to keep your 10 monetary units for yourself, the game in this round ends. In this case, your personal and the other person’s earnings in this round both equal 10 monetary units, i.e., the initial endowment.

**Transferring your 10 monetary units:** If you decide to transfer the 10 monetary units, we double this amount so that the other person receives 20 monetary units which are added to this person’s initial endowment. After you transfer your monetary units, the other person learns about your transfer and has to decide whether to transfer 10 monetary units back to you or to keep the monetary units she/he now possesses.

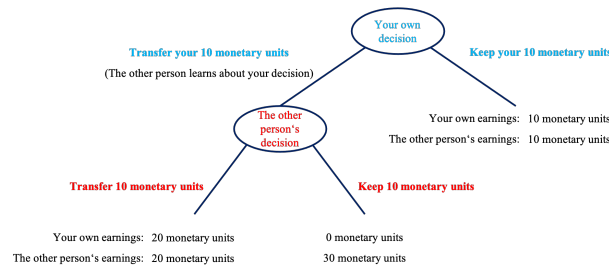
- If the other person transfers 10 monetary units back to you, we double this amount so that you receive 20 monetary units. In this case, your personal and the other person’s earnings in this round both equal 20 monetary units.
- If the other person does not transfer 10 monetary units back to you, your personal earnings equal 0 monetary units while the other person’s earnings equal 30 monetary units in this round.

Before you have to choose between transferring or keeping your 10 monetary units, you will receive information about 10 personal characteristics of the other person that is matched with you in a given round. The information might help you anticipate whether this other person will transfer 10 monetary units back to you so that you receive 20 monetary units, in case you initially decide to make a transfer. Note: The scale of characteristics that are not binary always reach from 'very low', 'low', 'medium', 'high', 'very high'.

Between rounds, you will not see the decision of the persons you are matched with. Part 1 ends once you have played 10 rounds of this game. We will then randomly select one of the rounds. The monetary units you own at the end of this round constitute your earnings for part 1. The other person matched to you in this round earns the number of monetary units she/he owns at the end of this round as well. Whether the earnings are payoff relevant for the other person is randomly determined. We will inform you about the decision of the other person, your earnings, and the other person's earnings from the selected round at the end of the experiment.

## Part 2

In part 2 of the experiment, you are randomly matched with another anonymous person from another study that you have not been matched with in part 1. You play one game that has the same structure as before:



Both you and the other participant receive 10 monetary units. Your task: You start making a decision about whether you want to keep your 10 monetary units or transfer all of them to the other person. Note: You can not transfer only a part of your endowment.

Keeping and transferring your monetary units has the same consequences as before:

**Keeping your 10 monetary units:** If you decide to keep the 10 monetary units for yourself, the game and part 2 end. In this case, your personal and the other person’s earnings in part 2 both equal 10 monetary units, i.e., the initial endowment.

**Transferring your 10 monetary units:** If you decide to transfer the 10 monetary units, we double this amount so that the other person receives 20 monetary units which are added to this person’s initial endowment. After you transfer your monetary units, the other person learns about your transfer and has to decide whether to transfer 10 monetary units back to you or to keep the monetary units she/he now possesses.

- If the other person transfers 10 monetary units back to you, we double this amount so that you receive 20 monetary units. In this case, your personal and the other person’s earnings in part 2 both equal 20 monetary units.
- If the other person does not transfer 10 monetary units back to you, your personal earnings equal 0 monetary units while the other person’s earnings equal 30 monetary units in part 2.

Before you have to choose between transferring or keeping your 10 monetary units, you will receive information about 3 out of 10 personal characteristics of the other person that is matched

with you. The information might help you anticipate whether the other person will transfer 10 monetary units back to you so that you receive 20 monetary units, in case you initially decide to make a transfer. Note: The scale of characteristics that are not binary always reach from 'very low', 'low', 'medium', 'high', 'very high'.

You decide which 3 characteristics of the other person you want to receive information about. You have to select one characteristic as the first choice, one characteristic as the second choice, and one characteristic as the third choice:

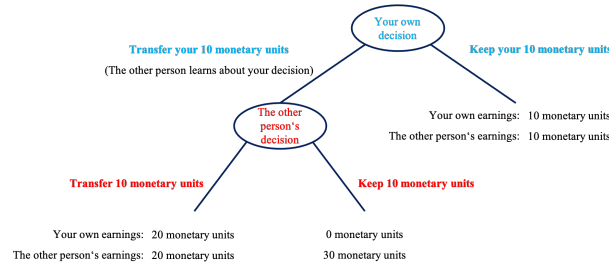
- **First choice:** The other person's characteristic you select as the first choice will be shown to you with a probability of 100%.
- **Second choice:** The other person's characteristic you select as the second choice will be shown to you with a probability of 90%. With a probability of 10% you will observe one of the other characteristics of this person that you neither selected as first, second, or third choice; which of these it is will be randomly determined.
- **Third choice:** The other person's characteristic you select as the third choice will be shown to you with a probability of 80%. With a probability of 20% you will observe one of the other characteristics of this person that you neither selected as first, second, or third choice and is not drawn before; which of these it is will be randomly determined.

After your selection decision, we will determine the three characteristics of the other person you will be able to see. You will then see the characteristics and be asked to make your transfer decision.

The monetary units you own at the end of the game constitute your earnings for part 2. The other person matched to you earns the number of monetary units she/he owns at the end of this part as well. Whether the earnings are payoff relevant for the other person is randomly determined. As before, you will not immediately learn about the decision of the other person. We will inform you about the decision of the other person, your earnings, and the other person's earnings from part 2 at the end of the experiment.

### Part 3

In part 3 of the experiment, you play 20 rounds of a game that has the same structure as in the previous parts of the experiment:



At the beginning of every round, you are randomly matched with a new anonymous person from another study. Both you and the other person receive 10 monetary units. Your task is always the same: You start making a decision about whether you want to keep your 10 monetary units or transfer all of them to the other person. Note: You can not transfer only a part of your endowment.

Keeping and transferring your monetary units has the following consequences:

**Keeping your 10 monetary units:** If you decide to keep the 10 monetary units for yourself, the game in this round ends. In this case, your personal and the other person's earnings in this round both equal 10 monetary units, i.e., the initial endowment.

**Transferring your 10 monetary units:** If you decide to transfer the 10 monetary units, we double this amount so that the other person receives 20 monetary units which are added to this person's initial endowment. After you transfer your monetary units, the other person learns about your transfer and has to decide whether to transfer 10 monetary units back to you or to keep the monetary units she/he now possesses.

- If the other person transfers 10 monetary units back to you, we double this amount so that you receive 20 monetary units. In this case, your personal and the other person's earnings in this round both equal 20 monetary units.
- If the other person does not transfer 10 monetary units back to you, your personal earnings equal 0 monetary units while the other person's earnings equal 30 monetary units in this round. Before you have to choose between transferring or keeping your 10 monetary units, you will receive information about 10 personal characteristics of the other person that is matched with you in a given round. These information might help you anticipate whether this other person will transfer

10 monetary units back to you so that you receive 20 monetary units, in case you initially decide to make a transfer. Note: The scale of characteristics that are not binary always reach from 'very low', 'low', 'medium', 'high', 'very high'.

In every round, a **Machine Learning System** produces a prediction about whether the person currently matched with you is most likely to transfer 10 monetary units back to you so that you receive 20 monetary units, if you initially decide to make a transfer. To make a prediction about a specific person, the Machine Learning System only uses the person's 10 personal characteristics that you also observe in the corresponding round.

The Machine Learning System is a Gradient Boosted Gradient Boosted Random Forest that was trained and tested on data from a previous study. Gradient Boosted Random Forest Classifier, despite their simplicity, are among the most powerful Machine Learning algorithms available today. They are widely used in a variety of domains by scientists and practitioners alike. In a test, the System used in the experiment reaches a recall score of 79.3%, which means that it correctly recognizes roughly 4 out of 5 people who actually reciprocate in case of a transfer. In other words, the Machine Learning System's prediction might help you better anticipate whether you will receive 20 monetary units in case you initially decide to make a transfer. Below you can find additional information about the structure of the system.

#### [ TREATMENT TEXT BEGIN

Together with the prediction you will receive an explanation about why the Machine Learning System makes a specific prediction about a person. For each the other person's 10 characteristics that the System uses to make the prediction the other person's individual characteristics that led to the prediction will be highlighted. More specifically, you will learn (i) the relative importance of the characteristic for the prediction about this specific person, and (ii) whether the specific characteristic (e.g. being female or male) contributes positively to the prediction that the person will return a transfer (in green) or is evidence against it (in red). Below you will find an example. In other words, for every prediction, you will receive an explanation why this prediction was made and which characteristics caused the prediction?

The importance of a characteristic and the direction of its impact are illustrated using colored bars.



- The relative length of a bar indicates the importance of a feature for the prediction. The longer the bar, the more pivotal is the characteristic for the specific prediction.
- A red bar indicates that the characteristic has a negative impact on the likelihood that the person returns monetary units back to you.
- A green bar indicates that the characteristic has a positive impact on the likelihood that the person returns monetary units back to you.

Example:



(i) The relatively long red bar indicates that, for this example, the Machine Learning System sees being highly competitive as relatively strong evidence against the person returning monetary units to you.

(ii) The relatively short green bar indicates that, for this example, the Machine Learning System sees having a high propensity to become upset as relatively weak evidence in favor of the person returning monetary units to you.

The technique to produce insights into why the Machine Learning System makes a specific prediction is called LIME (Local Interpretable Model-Agnostic Explanations). LIME attempts to understand black-box Machine Learning Systems by approximating individual predictions locally with an interpretable model. LIME was first introduced in 2016 by computer scientists from the University of Washington and has since become a state-of-the-art technique to render Machine Learning outputs transparent and interpretable. 'Explaining a prediction' refers to providing a human-interpretable understanding of the relationship between the inputs of a model (here the other person's 10 personal characteristics) and the model's prediction (whether the other person will reciprocate a transfer). For more information on LIME we refer the interested participant to the original research: (Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "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 (pp. 1135-1144).)

TREATMENT TEXT END ]

Gradient Boosted Random Forest, like its name implies, consists of a large number of individual decision Trees that operate as an ensemble. Based on examples, individual decision Trees learn logical rules which assign a certain label to new observations. These rules can be imagined as a sequence of consecutive questions. In the context of the game at hand, a single Tree could, for example, identify the following sequence of questions: 1. is the person open to new experiences? - Yes; 2. is the person female? - No; 3. is the person highly competitive? - Yes. Result: Given the answers to the question sequence, the person will most likely return monetary units back to you if you initially make a transfer. During the training process, the algorithm (more or less) automatically identifies the most informative questions to classify people as quickly as possible. Notably, each Tree in the Forest tries to correct the inaccuracies of previous Trees, thereby trying to boost the performance of the overall Forest. Gradient Boosted Random Forests typically comprise hundreds or even thousands of individual Trees.

Each individual Tree in a Gradient Boosted Random Forest spits out a prediction and the class (i.e. whether or not the other person returns 10 monetary units or not) with the most votes becomes the Gradient Boosted Random Forest's prediction. In other words: Knowing that individual Trees can be (randomly) wrong, we rely on the wisdom of the crowd, so that non-systematic errors of individual Trees cancel each other out. As a type of Ensemble Learner, Gradient Boosted Random Forests are among the most powerful Machine Learning algorithms currently available.

Between rounds, you will not see the decision of the persons you are matched with. part 3 ends once you have played 20 rounds of this game. We will then randomly select one of the rounds. The monetary units you own at the end of this round constitute your earnings for part 3. The other person matched to you in this round earns the number of monetary units she/he owns at the end of this round as well. Whether the earnings are payoff relevant for the other person is randomly determined. We will inform you about the decision of the other person, your earnings, and the other person's earnings from the selected round at the end of the experiment.

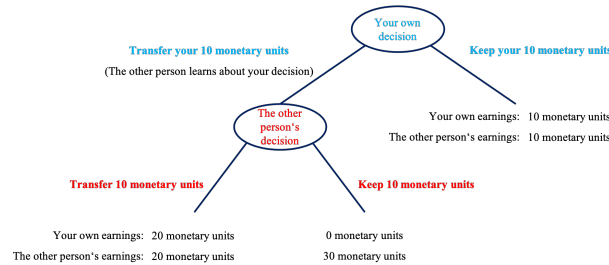
Now that you have finished all 20 rounds, we ask you to guess how good the Machine Learning System's predictions are. Overall you have to make three distinct guesses. For every guess that is not off by more than  $\pm 10$  percentage points from the actual value, you receive 5 monetary units.

Guess after 10 and 20 rounds: On a scale from 0% to 100% in steps of one percentage points, how often do you think the System produced a correct prediction for the first (second) 10 different persons you were matched with? A prediction was correct, whenever (i) the System predicted the

person to transfer 10 monetary units back to you, and the person would actually have done so if you had made a transfer, or (ii) the System predicted the person not to transfer 10 monetary units back to you, and the person would actually not have done so if you had made a transfer.

## Part 4

In part 4 of the experiment, you are randomly matched with another anonymous person from another study that you have not been matched with in any previous part. You play one game that has the same structure as before:



Both you and the other participant receive 10 monetary units. Your task: You start making a decision about whether you want to keep your 10 monetary units or transfer all of them to the other person. Note: You can not transfer only a part of your endowment.

Keeping and transferring your monetary units has the same consequences as before:

**Keeping your 10 monetary units:** If you decide to keep the 10 monetary units for yourself, the game and part 4 end. In this case, your personal and the other person's earnings in part 4 both equal 10 monetary units, i.e., the initial endowment.

**Transferring your 10 monetary units:** If you decide to transfer the 10 monetary units, we double this amount so that the other person receives 20 monetary units which are added to this person's initial endowment. After you transfer your monetary units, the other person learns about your transfer and has to decide whether to transfer 10 monetary units back to you or to keep the monetary units she/he now possesses.

- If the other person transfers 10 monetary units back to you, we double this amount so that you receive 20 monetary units. In this case, your personal and the other person's earnings in part 4 both equal 20 monetary units.
- If the other person does not transfer 10 monetary units back to you, your personal earnings equal 0 monetary units while the other person's earnings equal 30 monetary units in part 4.

Before you have to choose between transferring or keeping your 10 monetary units, you will receive information about 3 out of 10 personal characteristics of the other person that is matched

with you. The information might help you anticipate whether the other person will transfer 10 monetary units back to you so that you receive 20 monetary units, in case you initially decide to make a transfer. Note: The scale of characteristics that are not binary always reach from 'very low', 'low', 'medium', 'high', 'very high'.

You decide which 3 characteristics of the other person you want to receive information about. You have to select one characteristic as the first choice, one characteristic as the second choice, and one characteristic as the third choice:

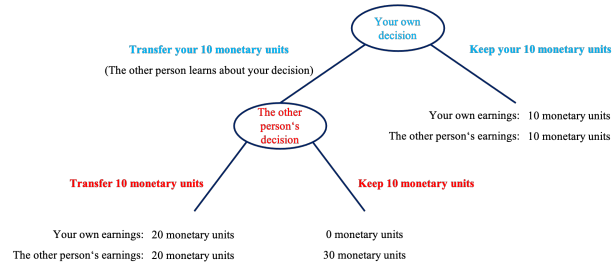
- **First choice:** The other person's characteristic you select as the first choice will be shown to you with a probability of 100%.
- **Second choice:** The other person's characteristic you select as the second choice will be shown to you with a probability of 90%. With a probability of 10% you will observe one of the other characteristics of this person that you neither selected as first, second, or third choice; which of these it is will be randomly determined.
- **Third choice:** The other person's characteristic you select as the third choice will be shown to you with a probability of 80%. With a probability of 20% you will observe one of the other characteristics of this person that you neither selected as first, second, or third choice and is not drawn before; which of these it is will be randomly determined.

After your selection decision, we will determine the three characteristics of the other person you will be able to see. You will then see the characteristics and be asked to make your transfer decision.

The monetary units you own at the end of the game constitute your earnings for part 4. The other person matched to you earns the number of monetary units she/he owns at the end of this part as well. Whether the earnings are payoff relevant for the other person is randomly determined. As before, you will not immediately learn about the decision of the other person. We will inform you about the decision of the other person, your earnings, and the other person's earnings from part 4 at the end of the experiment.

## Part 5

In part 5 of the experiment, you play rounds of a game that has the structure as before.



At the beginning of every round, you are randomly matched with a new anonymous person from another study. Both you and the other person receive 10 monetary units. Your task is always the same: You start making a decision about whether you want to keep your 10 monetary units or transfer all of them to the other person. Note: You can not transfer only a part of your endowment.

Keeping and transferring your monetary units has the following consequences:

**Keeping your 10 monetary units:** If you decide to keep your 10 monetary units for yourself, the game in this round ends. In this case, your personal and the other person's earnings in this round both equal 10 monetary units, i.e., the initial endowment.

**Transferring your 10 monetary units:** If you decide to transfer the 10 monetary units, we double this amount so that the other person receives 20 monetary units which are added to this person's initial endowment. After you transfer your monetary units, the other person learns about your transfer and has to decide whether to transfer 10 monetary units back to you or to keep the monetary units she/he now possesses.

- If the other person transfers 10 monetary units back to you, we double this amount so that you receive 20 monetary units. In this case, your personal and the other person's earnings in this round both equal 20 monetary units.

- If the other person does not transfer 10 monetary units back to you, your personal earnings equal 0 monetary units while the other person's earnings equal 30 monetary units in this round. Before you have to choose between transferring or keeping your 10 monetary units, you will receive information about 10 personal characteristics of the other person that is matched with you in a given round. The information might help you anticipate whether this other person will transfer 10 monetary units back to you so that you receive 20 monetary units, in case you initially decide

to make a transfer. Note: The scale of characteristics that are not binary always reach from 'very low', 'low', 'medium', 'high', 'very high'.

Between rounds, you will not see the decision of the persons you are matched with. Part 5 ends once you have played 10 rounds of this game. We will then randomly select one of the rounds. The monetary units you own at the end of this round constitute your earnings for part 5. The other person matched to you in this round earns the number of monetary units she/he owns at the end of this round as well. Whether the earnings are payoff relevant for the other person is randomly determined. We will inform you about the decision of the other person, your earnings, and the other person's earnings from the selected round at the end of the experiment.

### Questionnaire

The final part of the experiment consists of a questionnaire. Please read each question carefully and answer it truthfully. Once you have answered all questions, please press the "Next" button on your screen.

- What is your age?
- What is the highest academic degree you possess?
- What is your biological sex?
- How many years of working experience do you have?
- How would you classify the area you live in?
- Do you consider yourself more intelligent than the average person in the US?
- Do you consider yourself a better judge of character than the average person in the US?
- Do you consider yourself more talented than the average person in the US?
- I feel apprehensive about using technology:
- I have avoided technology because it is unfamiliar to me:

How well do the following statements describe you as a person? Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all" and a 10 means "describes me perfectly". You can also use any numbers between 0 and 10 to indicate where you fall on the scale, like 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

- When someone does me a favor I am willing to return it.
- If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so.
- I assume that people have only the best intentions.
- I enjoy being daring:

Please use a scale from 0 to 10, where 0 means you are "completely unwilling to take risks" and a 10 means you are "very willing to take risks". You can also use any numbers between 0 and 10 to indicate where you fall on the scale, like 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

- In general, how willing or unwilling you are to take risks.

Imagine the following situation: Today you unexpectedly received 1.000\$. How much of this amount would you donate to a good cause? (Values between 0 and 1.000 are allowed)? How much do you donate?

You are in an area you are not familiar with, and you realize that you lost your way. You ask a stranger for directions. The stranger offers to take you to your destination. Helping you costs the stranger about 20\$ in total. However, the stranger says he or she does not want any money from you. You have 6 presents with you. The cheapest present costs 5\$, the most expensive one costs 30\$. Do you give one of the presents to the stranger as a "thank-you"-gift? If so, which present do you give to the stranger? Please indicate the present you would give.

We will now use the computer to simulate the draw of a marble from a "cup". There are two cups, with different mixes of colored marbles, and you will be asked to guess the cup that is being used. First, we draw a computer-generated random number which will be either 1, 2, ... 6. Think of this as the throw of a die with 6 sides, with each side being equally likely.

- If the roll of the die yields 1 - 3, then the draw will be from the Green cup, which contains 2 green marbles and 1 yellow marble.
- If the roll of the die yields 4 - 6, then the draw will be from the Yellow cup, which contains 2 yellow marbles and 1 green marble.

You will not be told in advance the result of the die throw, so you will not know which cup is being used. Once the computerized die throw determines the cup to be used, you will be shown a randomly drawn marble from that cup.

You will get a chance to indicate the cup that you think is being used. Your money payoff will depend on whether your prediction turns out to be correct.

You will earn 2 monetary units for a correct prediction, and zero for an incorrect prediction. Considering the drawn marble, what cup do you think is it?



**Prior field study (for Study 1).** 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. We show the exact instructions of the field study in the following:

**How far do you live from your parents?**

Please select only one of the following answers:

- I live at my parents
- 1-10 KM away
- 11-50 KM away
- 51-150 KM away
- More than 150 KM away

**Have you, due to your studies, changed your place of residence?**

Please select only one of the following answers:

- Yes
- No

**How many siblings do you have?**

Please enter your answers below:

- Younger siblings [ ]
- Older siblings [ ]

**Please indicate with which hand you prefer to perform the following activities:**

(Always right, mostly right, both hands, mostly left, always left)

- Write [ ]

- Throw [ ]
- Tooth brushing [ ]
- Holding a spoon [ ]

**What languages do you speak at home? (multiple answers are possible)**

Please select all applicable answers:

- German
- Another language

**Please indicate with which hand you prefer to perform the following activities:**

(Mother and father)

- University
- University of applied science
- Technical college (former GDR)
- Technician or master craftsman examination
- Apprenticeship
- No educational background
- Unknown

**How do you finance yourself? (multiple answers are possible)**

(Please select all applicable answers:)

- My parents support me financially
- BAföG
- Scholarship
- Job as student assistant (Hiwi) at the university
- Job as a tutor at the university
- Job outside the university
- Other

**At which type of school did you get your university entrance qualification?**

(Please select only one of the following answers:)

- Grammar School
- Comprehensive school
- Vocational school
- Other

**After how many school years did you receive your university entrance qualification?**

(Please select only one of the following answers:)

- After less than 12 years
- After 12 years
- After 13 years
- After more than 13 years

**In which federal state of Germany did you acquire your university entrance qualification?**

(Please enter only one answers:)

[ ]

**Which of the following subjects did you take at school in the upper school and what grades (between 1.0 and 4.0) did you have in these subjects in your Abitur certificate?**

(Please select all applicable answers:)

- German
- English
- Physics
- Math

**Which of these subjects did you take as advanced courses at school?**

(Please select all applicable answers:)

- German
- English
- Physics
- Math
- None of these subjects

**On a scale from 1 (completely correct) to 6 (completely incorrect) please indicate the accuracy of the following statements.**

I chose my present course of study because...

- ...it particularly interested me and I wanted to.
- ...it corresponds to my inclinations and talents.
- ...as a graduate of this course of studies I expect particularly good earning and employment opportunities.
- ...I didn't know what else to do.

- ...I was influenced in my decision by my family / friends.

**Is your current course of study your dream study?**

(Please select only one of the following answers:)

- Yes
- No

**On a scale from 1 (completely sure) to 5 (completely unsure) please indicate the accuracy of the following statements.**

- How confident are you in your choice of study?
- How satisfied are you today with your choice of study?
- How certain are you that you will complete your studies?
- How certain are you that you will complete your studies at this University?

**Did you do one or more of the following activities before starting your current studies?**

Please select all applicable answers:

- Internship related to your field of study
- Internship not related to the field of study
- Training
- Completed studies
- Aborted studies
- Voluntary social year, German Armed Forces, Federal Voluntary Service etc.
- Other:

**How many semesters do you estimate you will need in total until you graduate from your current course?**

Please enter your answer below:

- Please enter your answer here [ ]

**What are your plans for the time after graduation from your current course of study?**

(Please select only one of the following answers:)

- Begin a further study (e.g. Master's degree)
- Start working

- Other

**Based on my grade point average, I expect to belong to...**

(Please select only one of the following answers:)

- the top [ ] percent of my year of study.

**How important is it to you to maintain your grade point average in your studies or even improve?**

(Please select only one of the following answers:)

- Very important
- Rather important
- Indifferent
- Rather unimportant
- Very unimportant

**How many hours a week do you think you should invest in your studies?**

Please enter your answer below:

- Please enter your answer here [ ]

**How many hours do you think you will actually invest in your studies each week?**

Please enter your answer below:

- Please enter your answer here [ ]

**How many hours a week do you currently invest in your studies?**

Please enter your answer below:

- Please enter your answer here [ ]

**Do you believe that your future earnings will depend on your final grade in your studies?**

Please select only one of the following answers:

- Completely applicable
- Mostly applicable
- Applies
- Mostly not applicable
- Completely not applicable

**How do you personally assess yourself? Are you generally a person willing to take risks or do you try to avoid risks?**

Please answer using the following scale, where the value 0 means: “Not willing to take risks at all”, and the value 10: “Very willing to take risks”. With the values in between you can grade your assessment.

- Please enter your answer here [ ]

**How do you personally assess yourself? Are you generally a person who is impatient or who is always very patient?**

Please answer using the following scale, where the value 0 means “very impatient” and the value 10 means “very patient”. With the values in between you can grade your assessment.

- Please enter your answer here [ ]

**To what extent do you agree with the following statement: “I’m a narcissist.” (Note: A narcissist is selfish, self-centered, vain.)?**

Please answer using the following scale, where a value of 1 means “do not agree at all” and a value of 5 means “agree completely”. With the values in between you can grade your assessment.

- Please enter your answer here [ ]

**How would you assess yourself in the context of the following statements?**

Please answer using the following scale, where 1 means “do not agree at all” and 5 means “agree completely”. The values in between allow you to grade your assessment.

- I like to find myself in situations where I am in competition with others.

**In the list below are different characteristics a person can have. It is likely that some characteristics will apply fully to you personally and others not at all. For others, you may be undecided.**

Please answer using the following scale from 1 to 5: A score of 1 means not applicable at all; 5 means fully applicable. With the values between 1 and 5 you can grade your opinion. I am someone who...

- works thoroughly
- is communicative, talkative
- is sometimes a little rough on others
- is original, brings in new ideas
- is forgiving
- is rather lazy

- can come out of herself/himself
- is sociable
- appreciates artistic, aesthetic experiences
- is easily nervous
- completes task effectively and efficiently
- is reserved
- is considerate and friendly with others
- has a vivid imagination
- is relaxed, can handle stress well

For the following decision situation, another survey participant will be assigned to you randomly. You and this other person make different decisions, which then result in your payout and the payout of the other person. At the beginning you and the other person will each receive 10 Euros from us. You have the following two options to choose from:

**Option A:** You keep your 10 Euros.

**Option B:** You give your 10 euros to the other person. The 10 Euros are doubled, i.e. the other person receives 20 Euros.

The other person also has these two options to choose from. Hence, there are four possible outcomes, depending on how you and the other person decide: If you and the other person both choose option A, you will both end up with 10 Euros each. If you and the other person both choose option B, both of you will each have 20 euros. If you choose option A and the other person chooses option B, you will have 30 euros and the other person 0 euros. And vice versa, if you choose option B and the other person chooses option A, you have 0 euros and the other person has 30 euros. In the following two situations, please decide whether you would rather choose option A or option B. The situations differ in whether you or the other person makes their decision first.

Situation 1: You decide first and the other person is informed of your decision. Which option do you choose?

- A / B

Situation 2: The other person makes their decision first, and you are informed of their decision. Which option do you choose if the other person has chosen option A?

- A / B

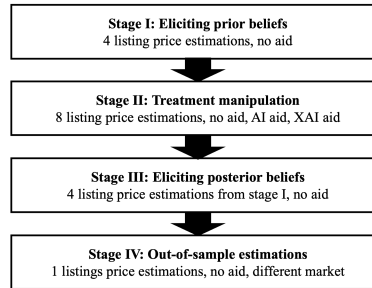
Which option do you choose if the other person has chosen option B?

- A / B



## Study 2: Experimental design

**Overview.** Our experiment takes place in the domain of real-estate where realtors, based on 10 observable apartment characteristics, need to predict property listing prices per square meters for several different objects. As AI-systems are increasingly available to produce first estimates of listing prices to support human evaluations, e.g., on Zillow.com in the US or Immowelt.de in Germany, this is a highly relevant domain.



**Figure 10** Sequence of Study 2

Notes: Sequence and overview of the four different stages in the experiment.

The experimental protocol comprises 4 stages (see Figure 10 for an overview). In Stage I, we elicit participants’ prior beliefs about the relation between apartment characteristics and the listing price. Stage II serves as our treatment manipulation. Conditional on the experimental condition they are in, participants make a series of listing price predictions without any aid, with the aid of an AI-system without explanations, or with the aid of an AI-system providing local explanations. In Stage III, we measure participants’ posterior beliefs of the relation between apartment characteristics and the listing price. Finally, in Stage IV participants make one final price prediction without any aid for a different apartment market. In the following subsections, we fill in the details of our experimental protocol.

**Details on listing price data and participant pool.** Throughout the study, participants, in one form or another, have to predict the listing price per square meter (hereafter listing price) for different apartments in German cities. To make their prediction, participants always observe ten features of the apartment. Importantly, the apartments participants encounter, differ only in regard to three out of ten features: whether or not the apartment has a balcony/terrace, the city where it is located, and the share of green voters in the city district (hereafter variable features). The other features are always fixed and identical across encountered listings (hereafter fixed features). The following Table 5 provides an overview of all apartment features. Always holding the same seven characteristics of an apartment fixed simplifies the price prediction task for participants,

Feature	Variability	Feature values
Apartment has balcony/terrace	Variable	Yes/No
Location	Variable	Frankfurt/Cologne (Chemnitz in Stage IV)
Share of green voters in district	Variable	Below city average / City average / Above city average
Year of construction	Fix	Between 2012 and 2022
Garden	Fix	No
Basement	Fix	Yes
Elevator in house	Fix	Yes
Floor	Fix	Second or third
Number of rooms	Fix	3
Unemployment numbers in district	Fix	Below city average / City average / Above city average

**Table 5** Used apartment features.

Notes: We show the apartment attributes that participants observed to make a decision.

mitigates potential concerns about information overload on the part of participants, and facilitates our analyses.

The listings participants encounter are real apartments that we scraped from a large German real-estate platform ([www.immonet.de](http://www.immonet.de)) over a period of 3 weeks in February 2022. The entire data set we collected comprises 5090 distinct observations. Excluding the observations that participants encounter in the study, we use the scraped data to develop an ML-based AI system that relies on the ten characteristics to predict listing prices. The ten characteristics include standard information available on the platform and additionally collected socio-economic information on the district where the apartment is located. The underlying ML model is a random forest whose hyperparameters we optimized via 5-fold cross-validation. The final model’s average  $R^2$  on a representative test set equals 72%. Importantly, in every experimental condition where participants interact with an AI system, the system’s overall listing price predictions for a given apartment are identical, i.e., originate from the exact same ML model. The explanations we provide in the corresponding treatment variations result from the post-hoc SHAP method (Lundberg and Lee 2017).

Our participant pool comprises experts from the real estate industry. More specifically, to recruit experts for our study, we collaborate with our industry partner Immobilienverband Deutschland (IVD). The IVD is a large German business association in the housing and real estate industry in the legal form of a registered association. Through our industry partner, we are able to contact approx. 6000 experts from the real estate industry in Germany which includes real estate agents, valuation experts, and property developers. We contact experts via the mail and invite them to take part in our study via a link. To ensure incentive compatibility and reduce attrition, we implement a contest incentive scheme. That is, we inform participants that for every correct listing price prediction they earn one point. A predicted listing price is correct if it does not differ from the scraped listing price by more than 500€. Participants only learn their overall score after finishing the entire experiment. After two weeks, we paid the ten participants with the highest scores 100€ each and issue an award-like certificate for their performance in accurately predicting listing prices.

If two participants earned the same number of points, we determine their ranking according to the sum of their predictions’ absolute deviation from the actual listing price.

**Stage I.** The purpose of Stage I is to elicit participants’ prior beliefs about the relationship between the three variable apartment characteristics and the objects’ actual listing price. To do so, we implement the following task. Participants encounter four apartments. As outlined above, the apartments differ only regarding having a balcony/terrace, location, and the share of green voters in the district, whereas all other features are fixed and identical. We randomly draw the four observed listings from the pool of the main examples ( $N=12$  given the permutations of variable features). For each listing participants encounter, they have to indicate marginal contributions of the given apartment’s variable features to its listing price. Participants can do so using a slider that ranges from minus to plus 2500€ in steps of 50€. As a reference point, we inform participants that the average listing price for an apartment that possesses seven fixed features is 9600€. By adjusting the three sliders whose default we set to 0€, participants change the overall estimated listing price for a given object whose default we set to the average of 9600€. For instance, assume that for a given apartment the values of the features Balcony/Terrace, Location, and Green voter share equal Cologne, Yes, Above average, respectively. If a participant sets the slider for Location to +1.000€, for Balcony/Terrace to -400€, and for Green voter share to 200€, the overall listing price prediction equals 11200€ ( $=9600+1000-400+200$ ). Additionally, we ask participants to state their confidence in their beliefs and the overall price prediction on a five-point scale. This procedure leaves us with point estimates for conditional prior beliefs (and confidence levels), which we can compare to identically elicited conditional posterior beliefs to identify adjustments on the individual level. Importantly, we randomize the draw of listings on the individual level so that we obtain the distribution of point estimates at the population level. Screenshots are provided in Figure 11 (original) and Figure 12 (English translation).

**Stage II.** In Stage II, participants have to predict the listing price for 8 listings. In contrast to Stage I, participants do not have to enter the contribution for the three variable apartment attributes. Instead, they only predict the overall listing price. We again ask participants to state their confidence in the price prediction on a five-point scale. Stage II introduces our treatment manipulations. We randomly assign participants to one of three different experimental conditions which differ in whether, and if so what type of AI-system support participants receive. In our baseline condition (NoAid), participants do not receive any support and make the price prediction entirely on their own. Participants in the AI condition observe the overall listing price prediction of the AI system, but do not obtain additional SHAP explanations about the system’s inner logic, i.e., they interact with a “black box” AI system. In our XAI condition, in addition to observing

## Eigentumswohnung 1/4

Klicken Sie auf "Fixe Eigenschaften", um sich erneut die identischen Eigenschaften der Immobilien anzusehen.

Fixe Eigenschaften

Variable Eigenschaften		Beitrag zum Preis/Quadratmeter
Stadt	Köln (Beitrag relativ zum Mittel der A-Städte in Deutschland)	<div> <div>0 EUR/qm</div> <div> <div></div> <div>-2500</div> <div>0</div> <div>2500</div> </div> </div> <p>Wie sicher sind Sie sich mit dieser Entscheidung? Angabe von 1 (unsicher) bis 5 (sicher).</p> <p><input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5</p>
Balkon/Terasse	Nein	<div> <div>0 EUR/qm</div> <div> <div></div> <div>-2500</div> <div>0</div> <div>2500</div> </div> </div> <p>Wie sicher sind Sie sich mit dieser Entscheidung? Angabe von 1 (unsicher) bis 5 (sicher).</p> <p><input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5</p>
Anteil Grünenwähler im Stadtteil (innerstädt. Vergleich)	Durchschnittlich	<div> <div>0 EUR/qm</div> <div> <div></div> <div>-2500</div> <div>0</div> <div>2500</div> </div> </div> <p>Wie sicher sind Sie sich mit dieser Entscheidung? Angabe von 1 (unsicher) bis 5 (sicher).</p> <p><input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5</p>

Ihre Preisschätzung: 9.600 EUR/qm

Wie sicher sind Sie sich mit dieser Entscheidung? Angabe von 1 (unsicher) bis 5 (sicher).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

(a) Original

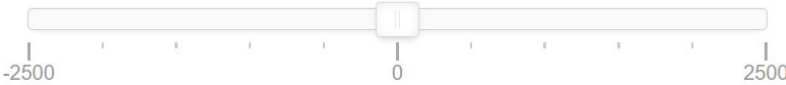


**Figure 11 Stage I and 3: Belief elicitation (Original).**

Notes: We show the original interface (in German) developed to let participants in Study 2 make listing price estimations in Stage I and 3. Participants entered their beliefs about the marginal contribution of apartment features to the overall listing price.

## Apartment 1/4

Click on "Fixed properties" to view the identical properties of the apartments again.

Fixed properties

Variable properties		Contribution to price/sqm
City	Cologne  (Contribution relative to the mean of the A-cities in Germany)	<div>0 EUR/sqm</div>  <p>How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).</p> <p><input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5</p>
Balcony/Terrace	No	<div>0 EUR/sqm</div>  <p>How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).</p> <p><input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5</p>
Proportion of Green voters in the district (inner-city comparison)	Average	<div>0 EUR/sqm</div>  <p>How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).</p> <p><input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5</p>

Your price estimate: 9.600 EUR/sqm

How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

(a) English translation

**Figure 12 Stage I and 3: Belief elicitation (English translation).**

Notes: We show the interface (English translation) developed to let participants in Study 2 make listing price estimations in Stage I and 3. Participants entered their beliefs about the marginal contribution of apartment features to the overall listing price.

the AI-system’s overall price prediction, participants also receive local SHAP explanations. More specifically, for every single listing they encounter, participants in the XAI condition observe the instance’s idiosyncratic SHAP values for the three variable apartment characteristics. We depict the local SHAP values right below the instance’s feature value. After they have finished all prediction tasks, participants in treatments with AI-system support (and explanations) fill out a survey containing items on their trust, degree of reliance, and perceived transparency of the AI-system (and explanations). These items serve as additional control variables in our analyses to detect potential treatment heterogeneities. Screenshots of the price prediction in the treatment stage are provided in Figure 13 (NoAid), in Figure 14 (AI), and in Figure 15 (XAI).

**Stage III.** In Stage III we again elicit participants’ beliefs about the relationship between the three variable apartment characteristics and the objects’ actual listing price, i.e., posteriors after making decisions (with the aid of an AI system) in Stage II. We elicit participants’ posterior beliefs simply by replicating Stage I, i.e., the same apartments. Note that independent of the treatment condition, participants do not receive any additional aid or information than they had previously. Again we also ask participants to state their confidence in their beliefs and overall listing price prediction. On an individual level, the measurement of posterior point estimates for beliefs, confidence levels, and importance levels allow us to observe adjustments per participant. A comparison of posterior distributions across different experimental variations further enables us to observe treatment effects on the population level distributions.

**Stage IV.** In Stage IV, we ask participants to make one final listing price prediction in the fashion of Stage II, i.e., provide an overall price prediction for a given listing and state the prediction confidence on a five-point scale. The seven fixed characteristics are again identical to all previously encountered apartments. The apartment is randomly drawn from a pool of instances with the same distribution of the Balcony and Green Voter characteristics, however, located in Chemnitz which is a mid-sized city in Eastern Germany. Participants do not obtain any additional aid. Given its location, we argue that the apartment is in a different apartment market (a mid-sized city in Eastern Germany). A comparison of aggregate distributions across treatments allows us to detect how belief adjustments affect listing prices in a different apartment market. Screenshots of this out-of-sample estimation are provided in Figure 16. After Stage IV, the study concludes with a brief questionnaire on participants’ socio-demographics including their age, gender, and working experience.

Eigentumswohnung 1/8

Klicken Sie auf "Fixe Eigenschaften", um sich erneut die identischen Eigenschaften und den Durchschnittspreis der Immobilien anzusehen.

Fixe Eigenschaften

Variable Eigenschaften	
Stadt	Frankfurt am Main
Balkon/Terasse	Nein
Anteil Grünenwähler im Stadtteil (innerstädt. Vergleich)	Unterdurchschnittlich

Ihre Preisschätzung in EUR/qm:

Wie sicher sind Sie sich mit dieser Entscheidung? Angabe von 1 (unsicher) bis 5 (sicher).

☐ 1

☐ 2

☐ 3

☐ 4

☐ 5

(a) Original

Apartment 1/8

Click on "Fixed Properties" to view again the identical properties and the average price of the apartments.

Fixed properties

Variable properties	
City	Frankfurt am Main
Balcony/Terrace	No
Proportion of Green voters in the district (inner-city comparison)	Below average

Your price estimate in EUR/sqm:

How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).

☐ 1

☐ 2

☐ 3

☐ 4

☐ 5

(b) English translation

Figure 13 Stage II: Treatment manipulation for NoAid condition.

Notes: We show the interfaces developed to let participants in Study 2 in the NoAid condition make listing price estimations in Stage II. For participants in this condition (NoAid), the interface shows only the fixed and variable characteristics of the apartment.

**Eigentumswohnung 1/8**

Klicken Sie auf "Fixe Eigenschaften", um sich erneut die identischen Eigenschaften und den Durchschnittspreis der Immobilien anzusehen.

Fixe Eigenschaften

Variable Eigenschaften	
Stadt	Frankfurt am Main
Balkon/Terasse	Nein
Anteil Grünenwähler im Stadtteil (innerstädt. Vergleich)	Unterdurchschnittlich

KI Vorhersage: 9.600 EUR/qm

Wie überrascht sind Sie von der Vorhersage der KI? Angabe von 1 (nicht überrascht) bis 5 (überrascht).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Ihre Preisschätzung in EUR/qm:

Wie sicher sind Sie sich mit dieser Entscheidung? Angabe von 1 (unsicher) bis 5 (sicher).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

(a) Original

**Apartment 1/8**

Click on "Fixed Properties" to view again the identical properties and the average price of the apartments.

Fixed properties

Variable properties	
City	Frankfurt am Main
Balcony/Terrace	No
Proportion of Green voters in the district (inner-city comparison)	Below average

AI Prediction: 9.600 EUR/sqm

How surprised are you by the AI's prediction? Indication from 1 (not surprised) to 5 (surprised).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Your price estimate in EUR/sqm:

How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

(b) English translation

**Figure 14 Stage II: Treatment manipulation for AI condition.**

Notes: We show the interfaces developed to let participants in Study 2 in the AI condition make listing price estimations in Stage II. For participants in this condition (AI), the interface shows the characteristics of the apartment and additionally the prediction of the AI system.



### Eigentumswohnung 1/8

Klicken Sie auf "Fixe Eigenschaften", um sich erneut die identischen Eigenschaften und den Durchschnittspreis der Immobilien anzusehen.

Fixe Eigenschaften

Variable Eigenschaften		KI Erklärung: Preisauswirkung
Stadt	Frankfurt am Main	+600 EUR/qm
Balkon/Terasse	Nein	-50 EUR/qm
Anteil Grünenwähler im Stadtteil (innerstädt. Vergleich)	Unterdurchschnittlich	-550 EUR/qm

KI Vorhersage: 9.600 EUR/qm

Wie überrascht sind Sie von der Vorhersage der KI? Angabe von 1 (nicht überrascht) bis 5 (überrascht).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Ihre Preisschätzung in EUR/qm:

Wie sicher sind Sie sich mit dieser Entscheidung? Angabe von 1 (unsicher) bis 5 (sicher).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

(a) Original

### Apartment 1/8

Click on "Fixed Properties" to view again the identical properties and the average price of the apartments.

Fixed properties

Variable properties		AI Explanation: Impact on price
City	Frankfurt am Main	+600 EUR/sqm
Balcony/Terrace	No	-50 EUR/sqm
Proportion of Green voters in the district (inner-city comparison)	Below average	-550 EUR/sqm

AI Prediction: 9.600 EUR/sqm

How surprised are you by the AI's prediction? Indication from 1 (not surprised) to 5 (surprised).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Your price estimate in EUR/sqm:

How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

(b) English translation

**Figure 15 Stage II: Treatment manipulation for XAI condition.**

Notes: We show the interfaces developed to let participants in Study 2 in the AI condition make listing price estimations in Stage II. For participants in this condition (XAI), the interface shows the characteristics, the AI prediction, and additionally SHAP values representing the impact of the three variable characteristics to the prediction (figures e/f).

## Eigentumswohnung

Klicken Sie auf "Fixe Eigenschaften", um sich erneut die fixen Eigenschaften der Immobilien anzusehen.

Fixe Eigenschaften

Variable Eigenschaften	
Stadt	Chemnitz (Sachsen)
Balkon/Terasse	Nein
Anteil Grünenwähler im Stadtteil (innerstädt. Vergleich)	Durchschnittlich

Ihre Preisschätzung in EUR/qm:

Wie sicher sind Sie sich mit dieser Entscheidung? Angabe von 1 (unsicher) bis 5 (sicher).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

(a) Original

## Apartment

Click on "Fixed Properties" to view again the identical properties and the average price of the apartments.

Fixed properties

Variable properties	
City	Chemnitz (Saxony)
Balcony/Terrace	No
Proportion of Green voters in the district (inner-city comparison)	Average

Your price estimate in EUR/sqm:

How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

(b) English translation

**Figure 16** Stage IV: Out-of-sample estimation

Notes: We show the interface developed to let participants in Study 2 make an out-of-sample listing price estimation in Stage IV. Panel (a) shows the original interface in German when participants entered their estimated listing price for an apartment in Chemnitz, Panel (b) shows the English translation.

## Instructions.

### Part 1

In part 1, your task is to estimate the price per square meter of four apartments offered on a real estate portal (i.e., it is the price called by the seller). For each estimate that differs from the real list price by no more than 500 EUR, you get one point. To help you make an informed decision, we always show you ten attributes of the apartments. Seven of the ten attributes are fixed (i.e. identical) for all apartments, so only three attributes vary.

[Table with 7 fixed properties]

Using sliders, you can specify the individual contribution of the three variable attributes to the offered price per square meter in euros. By adjusting the three sliders, you change the estimated price. As a reference and starting point, we show you the average price per square meter offered on the real estate portal for an apartment that has the seven fixed attributes and is located in a German “A-city”. A-cities are the seven most important German cities, namely Munich, Hamburg, Berlin, Stuttgart, Frankfurt am Main, Düsseldorf and Cologne.

[predicting prices of 4 apartments with sliders]

To conclude Part 1, we ask you to indicate how important you think the three variable attributes are for evaluating the price per square meter of the apartments. To do this, you can distribute 100 stars between the 3 attributes. The more stars you assign to a property, the more important you consider that property to be in evaluating the price. Please note that there are no right or wrong answers in these responses. We just want to better understand how you make the assessment.

[assigning attribute importance]

### Part 2

Your task now is to estimate the price per square meter offered on a real estate portal (i.e. the price called by the seller) for eight apartments. For each estimate that does not differ by more than 500€ from the real list price, you receive one point. Identical to Part 1, the apartments differ only in three out of ten attributes.

In contrast to part 1, you should now enter the offered price per square meter as a whole for each apartment. Again, we show you the average offered price per square meter of an apartment in a German A-city, which has the seven fixed attributes.

[BEGIN TEXT AI AND XAI]

As an aid to decision-making, this part provides you with the price prediction of an artificial intelligence (AI) previously developed by researchers at Goethe University. The AI was developed

to support real-estate experts in their valuation decisions. Note that you are not bound by the prediction.

The AI uses the ten displayed attributes of apartments to predict the price per square meter offered. The AI is based on a Random Forest, one of the simplest but also one of the most powerful AI methods. A Random Forest uses a large number of different decision trees, each predicting a single value (in this case, the price/sqm). The majority prediction of all decision trees then determines the final prediction. In other words, the Random Forest uses the “wisdom of crowds.”

Several performance metrics show that the AI trained for this study is good at predicting the offered price per square meter of apartments. In one test, the AI was able to explain over 70 % of the variation in price per square meter. Thus, the AI can potentially help you make an accurate valuation.

END TEXT NoAid AND XAI]

[BEGIN TEXT XAI

In addition to the AI’s prediction, you will receive explanations on how the AI arrives at individual price predictions for specific apartments. For this purpose, the AI system explains to you the individual contribution of the three variable attributes to the prediction of the price per square meter of individual apartments in German A-cities. These explanations should help to make the behavior of the AI transparent and interpretable. You will find the individual contributions next to each of the variable attributes.

END TEXT XAI]

[predicting prices for 8 apartments directly]

### **Part 2 questionnaire (AI and XAI only)**

To conclude Part 2, we ask you to answer a few questions about the AI truthfully. Please note that there is no right or wrong in these answers. We just want to better understand how you approach the evaluation.

[BEGIN QUESTIONS AI AND XAI

Please indicate your agreement with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree).

- I include AI’s advice in my evaluation of the price per square meter.

On a scale from 0 to 100%:

- How accurate do you think the AI's price predictions are?

Please indicate your agreement with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree).

- The AI is competent and effective in predicting the listed price per square meter.
- The AI does a very good job at predicting the listed price per square meter.
- Overall, the AI is a competent help for my evaluation of the price per square meter.

Please indicate your agreement with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree).

- The AI gives unbiased assessments.
- The AI is honest.
- I consider this AI to have integrity.

Please indicate your agreement with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree).

- I feel safe relying on the AI to make my decision.
- I feel comfortable relying on the AI to make my decision.
- I feel satisfied when I rely on the AI to make my decision.

END QUESTIONS AI AND XAI]

[BEGIN QUESTIONS XAI

Please indicate your agreement with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree).

- I include explanations in my evaluation of the price per square meter.

On a scale from 0 to 100%:

- How accurate do you think the AI's explanations are?

Please indicate your agreement with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree).

- The explanations are competent and effective at conveying the logic of AI.
- The explanations do your job of conveying the logic of AI very well.
- Overall, the explanations are a competent help to understand the logic of AI.

Please indicate your agreement with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree).

- The explanations are unbiased.
- The explanations are honest.
- I consider the explanations to have integrity.

Please indicate your agreement with the following statements on a scale of 1 (strongly disagree) to 7 (strongly agree).

- I feel safe relying on the explanations to make my decision.
- I feel comfortable relying on the explanations to make my decision.
- I feel satisfied when I rely on the explanations to make my decision.

END QUESTIONS XAI]

### Part 3

In Part 3, your task is to estimate the price per square meter offered on a real estate portal (i.e., it is the price called by the seller) for four apartments.

[Table with 7 fixed properties]

Using sliders, you can specify the individual contribution of the three variable attributes to the offered price per square meter in euros. By adjusting the three sliders, you change the estimated price. As a reference and starting point, we show you the average price per square meter offered on the real estate portal for an apartment that has the seven fixed attributes and is located in a German “A-city”. A-cities are the seven most important German cities, namely Munich, Hamburg, Berlin, Stuttgart, Frankfurt am Main, Dusseldorf and Cologne.

[predicting prices of 4 apartments with sliders]

To conclude Part 3, we again ask you to indicate how important you think the three variable attributes are in evaluating the price per square meter of apartments. To do this, you can distribute 100 stars between the 3 attributes. The more stars you assign to a property, the more important you consider that property to be in evaluating the price. Please note that there are no right or wrong answers in these responses. We just want to better understand how you make the assessment.

[assigning attribute importance]

### Part 4

In the last part of this study, your task is to estimate the offered price per square meter (so it is the price called by the seller) for one final apartment. If estimate that does not differ from the real list price by more than 500€, you will receive one point. As before, we show you ten attributes of the apartment, with the fixed seven apartments identical to the previous apartments.

Analogous to part two, you should enter the offered price per square meter for the two apartments.

[predict prices of Chemnitz apartment directly]

## Questionnaire

To complete the study, we ask you to truthfully fill out a short questionnaire.

- How old are you?
- What is your gender?
- What is your highest academic degree?
- How many years of professional experience in the real estate industry do you have?

On a scale of 0 (not at all) to 10 (extremely much):

- How much experience in the valuation of apartments do you have?

On a scale of 1 (strongly disagree) to 7 (strongly agree):

- I think I'm better at accurately valuing real estate properties than the average real-estate expert in Germany.

- I think that I am smarter than the average German.

On a scale from 0 (not at all) to 10 (extremely much):

- In general, how willing are you to take risks?
- I am familiar with predictive software that provides information to support human decision-making.

Your e-mail address [ ]

**Information on the dataset and AI system.** We obtained the dataset by crawling apartments listed on large online platform in February 2022. Specifically, we considered apartments listed for sale in the seven major cities of Germany (“A-cities”) and scraped multiple different attributes reflecting the number of rooms in the apartment or whether it has a balcony. We disregarded apartments for which the information on one or several attributes was missing. In order to characterize the location of the apartment within the city, we joined third party data from public statistics: the share of voters for the German green party and the unemployment rate. Both attributes are captured on the level of districts and, subsequently, bagged to lower, mid, and upper third within the respective city. For example, if an apartment in Berlin is in the low third for unemployment, then it is located in a district for which the unemployment rate is below the average unemployment rate in Berlin. We further treat the top 0.5% of apartments with regard to the listing price as outliers and exclude them from our data. The final, preprocessed dataset comprises 5090 apartments and is described in Table 6.

Continuous attributes	average	standard dev	0.25 quantile	median	0.75 quantile
Listing price/ $m^2$ [€]:	7158.55	3217.37	4500.0	6500.0	8500.0
Construction [year]:	1971.18	43.07	1937.0	1972.0	2018.0
Nmbr of rooms:	2.72	1.25	2.0	3.0	3.0
Floor (storey):	1.80	2.56	0.0	1.0	3.0
Ordinal attributes			lower third	mid third	higher third
Unemployment			44.7 %	30.8 %	24.6 %
Green party electorate			39.1 %	25.8 %	35.1 %
Binary attributes			Yes		No
Basement			68.1 %		31.9 %
Elevator			45.3 %		54.7 %
Balcony			60.1 %		39.9 %
Garden			21.5 %		78.5 %
Multicat. attributes			Distribution (shares)		
City			Berlin (39.2 %), Hamburg (19.4 %), Munich (16.1 %) Cologne (8.9 %), Frankfurt (7.0 %), Stuttgart (4.8 %) Dusseldorf (4.7 %)		

**Table 6** Descriptive statistics of real-estate data.

Notes: We scraped the data from a large real-estate platform in Germany and joined the ordinal attributes (unemployment and green party electorate) by drawing from public statistics. We considered the seven major cities in Germany (“A-Cities”, the east German city of Chemnitz is not included here). We excluded real-estate for which the price or any of the remaining attributes were not listed. This left us with 5090 observations.

We randomly split the data into different sets for training (95%) and testing (5%) of our AI system, following common conventions. Moreover, we ensure that the apartments directly featured in our experiment fall into the test set.

Our AI system is based on a random forest. To yield a prediction, the random forest averages across the predictions of multiple, randomized decision trees. In our case, the random forest predicts the listing price per square meter based on the remaining 10 attributes as predictors. We determine the hyperparameters for the random forest by applying a grid search in a 5-fold cross-validation



on the training set. Subsequently, we assess the performance of our AI system based on the test data ( $R^2 = 0.72$ ).

Our explanations are based on SHAP values. We compute SHAP values for all predictors using the tree implementation of the SHAP value method. As a result, for each of the 12 apartments featured in the experimental Stages I to III, we yield both the predicted listing price per square meter and the contribution of each of the 10 predictors.

## Study 1: Analyses

**Relationship between LIME and feature values.** Table 7 provides information about the relationship between borrower characteristics and associated LIME values. We depict the estimated coefficient and the adjusted  $R^2$  resulting from simple OLS regressions where the trait serves as the dependent variable and the LIME value is the only independent variable. We also report p-values for the estimated coefficients.

Attribute	Coefficient	Adj. $R^2$
Competit.	-0.91, $p < 0.01$	0.81
Openness	0.36, $p < 0.01$	0.12
Conscient.	-0.13, $p = 0.37$	0.00
Agreeabln.	-0.07, $p = 0.61$	0.00
Neuro.	0.85, $p < 0.01$	0.70
Extrav.	0.78, $p < 0.01$	0.63
Patience	0.68, $p < 0.01$	0.45
Younger sibl.	-0.85, $p < 0.01$	0.73
Older sibl.	-0.66, $p < 0.01$	0.41
Gender	-0.98, $p < 0.01$	0.96

**Table 7** Multicollinearity between characteristics and LIME.

Notes: We depict coefficients, associated p-values, and adjusted  $R^2$  measures from OLS regressions, where LIME values for a borrower trait serve as the only independent and the actual borrower traits as dependent variables. Reported results provide insights into the multicollinearity between LIME and trait values.

For most borrower traits, there is a strong relationship between their actual value and the associated LIME value. This relationship manifests in coefficient estimates depicting high, almost perfect correlations and high adjusted  $R^2$  values revealing strong explanatory power for the variation in the characteristic. Hence, using borrower characteristics and associated LIME values simultaneously as independent variables in regression analyses creates multicollinearity problems (e.g., measured by the Variance Inflation Factor). Depicted values reveal that only for *Openness*, *Conscientiousness*, and *Agreeableness* the correlation seems somewhat contained.

**Parallel trends assumption.** Table 8 reports regression results where participants' investment decisions in Stage I serve as the dependent and observed borrower traits as the independent variables. Columns (1) and (2) show estimates for baseline and treatment participants, respectively. Column (3) reports estimates for treatment differences between estimates reported in columns (1) and (2), i.e., coefficients for borrower trait and treatment interaction terms in a pooled regression. In all three models, we include individual fixed effects and report robust standard errors in parentheses. Reported estimates are standardized to facilitate comparability.

Depicted regression results provide insights into the validity of interpreting Difference-in-Difference estimates for distinct borrower traits. Put differently, the analyses in Table 8 test the parallel trends assumption. According to our regression results, there are no statistically significant

Dep. variable:	(1)	(2)	(3)
Investing in Stage I	Baseline (AI)	Treatment (XAI)	$\Delta(1) - (2)$
Competit.	-0.038*** (0.011)	-0.037*** (0.010)	0.001 (0.015)
Openness	0.025*** (0.008)	0.027*** (0.008)	0.002 (0.012)
Conscien.	0.004 (0.008)	0.018** (0.008)	0.014 (0.012)
Agreeabln.	0.082*** (0.010)	0.083*** (0.010)	0.000 (0.014)
Neuroticism	-0.031*** (0.010)	-0.009 (0.010)	0.022 (0.015)
Extrav.	0.028*** (0.009)	0.033*** (0.009)	0.006 (0.013)
Patience	0.031*** (0.008)	0.037*** (0.008)	0.006 (0.011)
Younger sibl.	-0.005 (0.009)	-0.018** (0.008)	-0.013 (0.012)
Older sibl.	0.027*** (0.008)	0.025*** (0.008)	-0.001 (0.012)
Gender (Male)	-0.038*** (0.009)	-0.015 (0.010)	0.023 (0.014)
N	3060	3010	6070
p	0.000	0.000	0.000
R <sup>2</sup>	0.385	0.446	0.416

**Table 8** Check for parallel trends assumption.

Notes: We depict results for OLS regressions with fixed effects. We report robust standard errors in parentheses. Participants' investment decisions in Stage I serve as the dependent variable. As independent variables, we include all borrower traits and the borrower's actual type. Column (1) shows results for baseline (AI) participants, column (2) shows results for treatment (XAI) participants, and column (3) shows estimated differences between coefficients in columns (1) and (2). Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

treatment difference in participants' initial weighting of borrower traits (see column (3)). However, looking at the magnitude and significance of estimates in columns (1) and (2) together, we find that only in one of the two conditions do participants consider *Conscientiousness*, *Neuroticism*, *Younger Siblings*, and *gender*. Hence, despite the statistical insignificance of these estimated treatment differences, there is reason to believe that there is a relevant difference, calling into question the interpretation of corresponding DiD estimates. Against this background, we will refrain from interpreting these DiD estimates.

**Situational information processing.** Table 9 reports results from fixed-effects OLS regression according to model (1), setting  $s = 2$ . Different columns show results for different subsamples of our data. Using different subsamples renders some of the dummy variables in model (1) constant, effectively reducing the model. Columns (1) and (2) show  $\beta_1$  estimates for baseline participants' decisions in Stages I and II, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows  $\beta_2$  estimates for baseline participants, measuring weight changes driven by the provision of explanations. Finally, column (6) shows DiD estimates  $\beta_4$ , i.e., isolated explanation-driven weight changes.

Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage II	Stage I	Stage II	Effect	Effect
Competit.	-0.039*** (0.011)	-0.020*** (0.007)	-0.035*** (0.011)	-0.096*** (0.007)	0.019 (0.012)	-0.084*** (0.017)
Openness	0.024*** (0.008)	0.011** (0.005)	0.029*** (0.008)	0.013* (0.007)	-0.013 (0.009)	-0.004 (0.014)
Agreeabln.	0.081*** (0.010)	0.049*** (0.007)	0.085*** (0.010)	0.009 (0.007)	-0.032*** (0.012)	-0.038** (0.017)
Extrav.	0.027*** (0.009)	0.012** (0.006)	0.034*** (0.009)	0.002 (0.011)	-0.016 (0.011)	-0.006 (0.017)
Patience	0.031*** (0.008)	0.005 (0.006)	0.037*** (0.008)	0.029*** (0.009)	-0.026*** (0.009)	0.035** (0.014)
Older sibl.	0.028*** (0.009)	0.000 (0.005)	0.023*** (0.008)	0.025*** (0.009)	-0.028*** (0.010)	0.020 (0.014)
Repayment pred.		0.224*** (0.012)		0.164*** (0.008)		-0.051*** (0.016)
N	3060	6120	3010	6020	9180	18210
p	0.000	0.000	0.000	0.000	0.000	0.000
R <sup>2</sup>	0.385	0.453	0.446	0.410	0.386	0.430

**Table 9** Change in information weighting across Stages I and II.

Notes: We depict results for OLS regressions with fixed effects. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and II serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and II, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Depicted regression results provide insights into the effects of providing predictions and explanations on situational information processing. The estimates form the basis of our Figure 2.

We mainly rely on fixed-effect OLS models instead of non-linear models such as logit or probit for our analyses. That is because our main interest lies in interaction terms capturing the isolated effects of observing predictions and explanations, i.e., cross-partial derivatives. For non-linear models like logit or probit, marginal effects are not constant over their range. As a consequence, the statistical significance of interaction term coefficients cannot be tested with simple asymptotic z-statistics. In addition to this limitation, the sign of interaction term coefficients not necessarily indicates the direction of the cross-partial effect (see, e.g., Ai and Norton 2003). Given the variation in the ten borrower traits and different interaction levels, there is a valid concern that estimates for non-linear models provide inappropriate insights into the existing effects. Notably, despite the possible pitfalls in using non-linear models, the estimated marginal effects for OLS models and estimated marginal effects at the mean for logit models are convincingly similar in direction and significance so we are confident that our results are not an artifact of our model selection. We rerun models depicted in Table 9 using a logit model to demonstrate the similarity. Table 10 shows estimated marginal effects at the mean. A comparison of Tables 9 and 10 reveals that the estimates are almost identical.

Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage II	Stage I	Stage II	Effect	Effect
Competit.	-0.072*** (0.019)	-0.045*** (0.014)	-0.072*** (0.021)	-0.156*** (0.016)	0.031 (0.021)	-0.116*** (0.034)
Openness	0.036** (0.014)	0.022** (0.011)	0.048*** (0.016)	0.026** (0.011)	-0.017 (0.017)	0.002 (0.028)
Agreeabln.	0.127*** (0.017)	0.097*** (0.014)	0.149*** (0.020)	0.011 (0.012)	-0.035* (0.020)	-0.094*** (0.033)
Extrav.	0.047*** (0.014)	0.020* (0.012)	0.067*** (0.016)	0.035** (0.016)	-0.031* (0.019)	0.058 (0.041)
Patience	0.054*** (0.013)	0.015 (0.012)	0.075*** (0.015)	0.082*** (0.013)	-0.037** (0.016)	0.044* (0.026)
Older sibl.	0.041*** (0.013)	-0.001 (0.009)	0.045*** (0.015)	0.020** (0.009)	-0.046*** (0.017)	0.010 (0.026)
Repayment pred.		0.311*** (0.019)		0.243*** (0.017)		-0.074** (0.036)
Observations	2380	5580	2240	5580	7960	15780
p	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo $R^2$	0.134	0.353	0.17	0.291	0.294	0.279

**Table 10** Change in information weighting across Stages I and II – Logit.

Notes: We depict results for logit regressions with fixed effects. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and II serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and II, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

**Influence of LIME values on decision making.** Table 11 shows regression results where participants’ investment decisions in Stage II serve as the dependent variable. As independent variables, we include the LIME values, the observed prediction, and the borrowers’ *Openness*, *Conscientiousness*, and *Agreeableness* (we exclude the others due to the aforementioned multicollinearity problems), and these LIME values and feature values interaction with a treatment dummy. We further include individual fixed effects and report robust standard errors in parentheses. Reported estimates are standardized to facilitate comparability. Importantly, we report the estimates for  $\text{LIME} \times \text{Treatment}$  interaction effects. The reason is that LIME values are strongly related to predictions so we find significant LIME effects even for baseline participants, who did not observe them in Stage II. By looking at the additional effect that the actually observing LIME values have, we are able to draw appropriate conclusions about their influence on participants’ investment decisions.

Dep. variable:	(1)
Investing in Stage II	
LIME Competit.	0.088*** (0.012)
LIME Openness	0.006 (0.008)
LIME Agreeabln.	0.016** (0.008)
LIME Extrav.	0.010 (0.009)
LIME Patience	0.026*** (0.008)
LIME Older sibl.	0.005 (0.007)
N	12140
p	0.000
$R^2$	0.435

**Table 11** Relationship between LIME values and investments for treatment participants.

Notes: We depict results for OLS regressions with fixed effects. Participants’ investment decisions in Stages II serve as the dependent variable. As independent variables, we include all LIME values, borrower traits that do not create multicollinearity, observed predictions, the borrower’s actual type, and interaction effects for these variables with a treatment dummy. Reported estimates represent  $\text{LIME} \times \text{Treatment}$  dummy interaction terms so that we can control for correlations between predictions and LIME values. Estimates are standardized. We report robust standard errors in parentheses. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Depicted results provide direct insights into whether, and if so how, participants’ investment decisions depend on the actually observed LIME values. Two results are important to our argumentation. First, we find that all estimates for LIME values are positive, indicating that treatment participants’ investment decisions do indeed vary with the observed LIME values. For instance, participants are *ceteris paribus* more (less) likely to invest when observing positive (negative) LIME values for competitiveness. Second, we find that only the two highest LIME values (for *Competitiveness* and *Patience*) and the trait participants initially put most emphasis

on (*Agreeableness*) are statistically significant. Hence, participants do not seem to consider all explanations equally but only look at some of them more closely. Notably, the LIME values they put the most weight on belong to the traits for which we observe significant explanation effects.

**Mental model adjustments.** Table 12 reports results from fixed-effects regression according to model (1), setting  $s = 3$ . Different columns show results for different subsamples of our data.

Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
	Stage I	Stage III	Stage I	Stage III	Effect	Effect
Investing						
Competit.	-0.039*** (0.011)	-0.044*** (0.011)	-0.035*** (0.011)	-0.087*** (0.013)	-0.005 (0.013)	-0.048** (0.019)
Openness	0.024*** (0.008)	0.026*** (0.009)	0.029*** (0.008)	-0.001 (0.009)	0.001 (0.011)	-0.031** (0.015)
Agreeabln.	0.081*** (0.010)	0.097*** (0.011)	0.085*** (0.010)	0.078*** (0.010)	0.016 (0.011)	-0.023 (0.016)
Extrav.	0.027*** (0.009)	0.045*** (0.010)	0.034*** (0.009)	0.024** (0.010)	0.018* (0.011)	-0.027* (0.016)
Patience	0.031*** (0.008)	0.016* (0.009)	0.037*** (0.008)	0.059*** (0.010)	-0.015 (0.010)	0.036** (0.015)
Older sibl.	0.028*** (0.009)	0.033*** (0.009)	0.023*** (0.008)	0.018** (0.009)	0.005 (0.011)	-0.010 (0.015)
N	3060	3060	3010	3010	9180	12140
p	0.000	0.000	0.000	0.000	0.000	0.000
R <sup>2</sup>	0.385	0.387	0.446	0.393	0.386	0.404

**Table 12** Change in information weighting across Stages I and III.

Notes: We depict results for OLS regressions with fixed effects. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and III serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and III, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Using different subsamples renders some of the dummy variables in model (1) constant, effectively reducing the model. Columns (1) and (2) show  $\beta_1$  estimates for baseline participants' decisions in Stages I and III, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows  $\beta_2$  estimates for baseline participants, measuring weight changes driven by the provision of explanations. Finally, column (6) shows DiD estimates  $\beta_4$ , i.e., isolated explanation-driven weight changes.

Depicted regression results provide insights into the effects of providing predictions and explanations on mental model adjustment processes. The estimates form the basis of our Figure 3.

**Investment decision performance.** Table 13 reports regression results for different models in which either the accuracy or recall measures serve as the dependent variable. Our independent variables of main interest are the treatment dummy XAI, a dummy for borrowers with the highest competitive levels, and the interaction of these two dummies. We additionally control for observed

borrower traits, and, for regressions in columns (2) and (5), the observed prediction and LIME values. We report robust standard errors in parentheses.

	Accuracy			Recall		
	(1)	(2)	(3)	(4)	(5)	(6)
	Stage I	Stage II	Stage III	Stage I	Stage II	Stage III
XAI ( $\alpha_1$ )	0.006 (0.020)	-0.013 (0.016)	-0.032 (0.020)	0.033 (0.027)	-0.015 (0.020)	-0.016 (0.026)
Very high Competit. ( $\alpha_2$ )	-0.082*** (0.027)	-0.037** (0.017)	-0.112*** (0.028)	0.011 (0.035)	-0.009 (0.021)	0.021 (0.037)
XAI $\times$ Very high Competit. ( $\alpha_3$ )	-0.004 (0.026)	-0.109*** (0.019)	-0.023 (0.026)	-0.024 (0.031)	-0.149*** (0.024)	-0.089*** (0.033)
F-test: $\alpha_1 + \alpha_3 = 0$	$p = 0.91$	$p < 0.01$	$p < 0.03$	$p = 0.77$	$p < 0.01$	$p < 0.01$
N	6070	12140	6070	4697	9752	4697
p	0.000	0.000	0.000	0.000	0.000	0.000
$R^2$	0.023	0.072	0.022	0.046	0.139	0.066

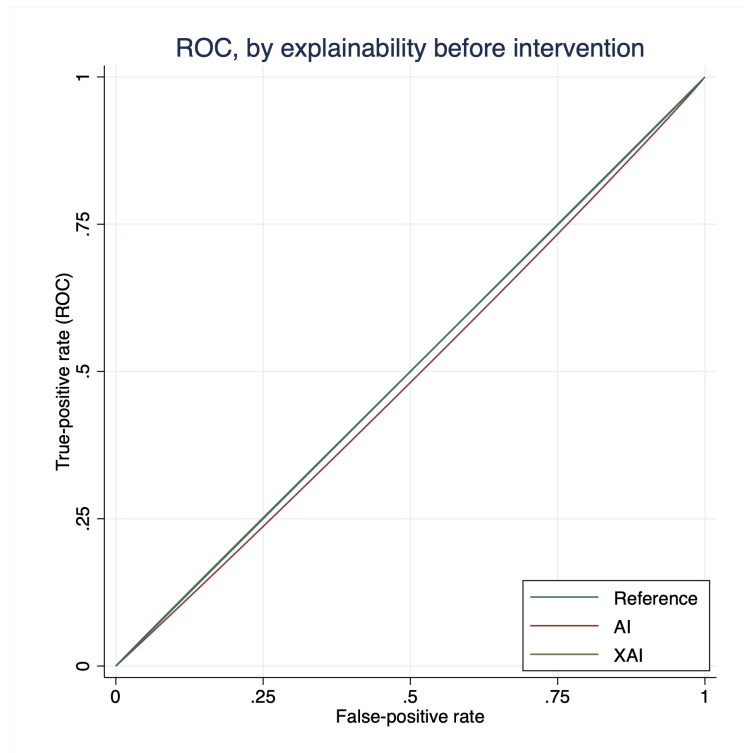
**Table 13** Treatment differences for different performance measures.

Notes: We depict results for OLS regressions. We report robust standard errors in parentheses. In columns (1) to (3), we use a dummy as the dependent variable that indicates whether a participant made the payoff maximizing investment decision – Accuracy. In columns (4) to (6), we use a dummy as the dependent variable that indicates whether a participant correctly invested with a repaying borrower – Recall. The independent variables of main interest are a treatment dummy, a dummy indicating that the borrower is most competitive, and their interaction term. We additionally control for borrowers’ other traits, and, if appropriate, for the observed prediction and LIME values. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Depicted estimates reveal that the treatment differences in participants’ accuracy and recall in decision-making stem from instances where borrowers are most competitive. In Stages II and III we do not find that observing explanations does generally decrease the investment performance. Instead, treatment differences only occur for borrowers with the highest levels of competitiveness. Importantly, participants observe the most negative LIME values (also highest in absolute terms) for this level of Competitiveness. It, therefore, seems that observing these highly negative LIME values leads participants to make worse decisions.

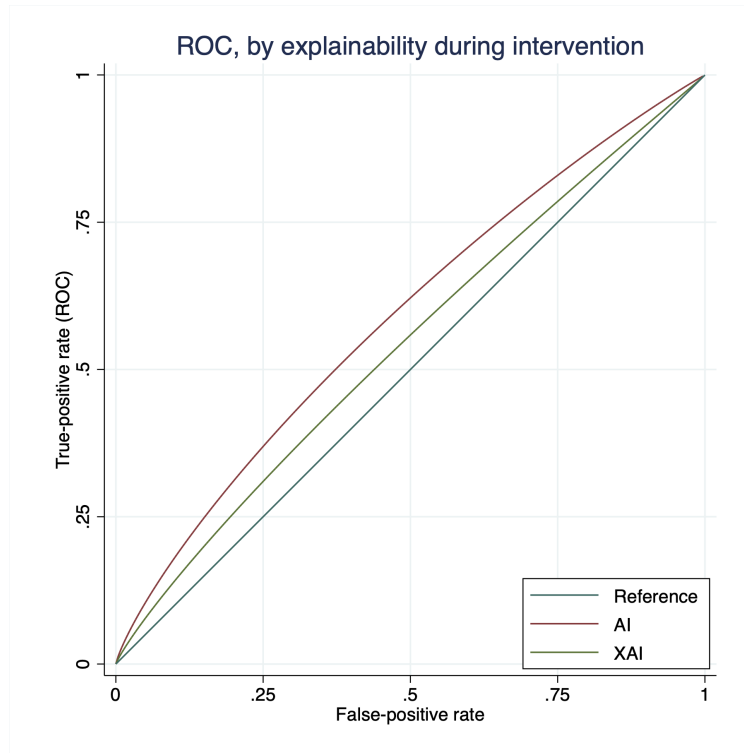
In the following, we depict ROC curves that provide insights into the optimality of participants’ investment decisions. To construct the plots, the borrowers’ actual repayment behavior serves as the actual class (1=Repayment, 0=No repayment), whereas participants investment decisions serve as the predicted class (1=Making an investment, 0=Not making an investment). Importantly, neither of these plots depicts the pure performance of the (X)AI system’s prediction. For Stage II, where participants interacted with the system and observed predictions, the corresponding ROC curve depicts the performance of participants’ final decision that may or may not be affected by the observed prediction (and explanations). We depict the performance of the underlying system alone in Stage II in Figure 20. We show images separately for participants’ decisions before, during, and after the treatment intervention. In each Figure, we show ROC plots for the AI and the XAI conditions.





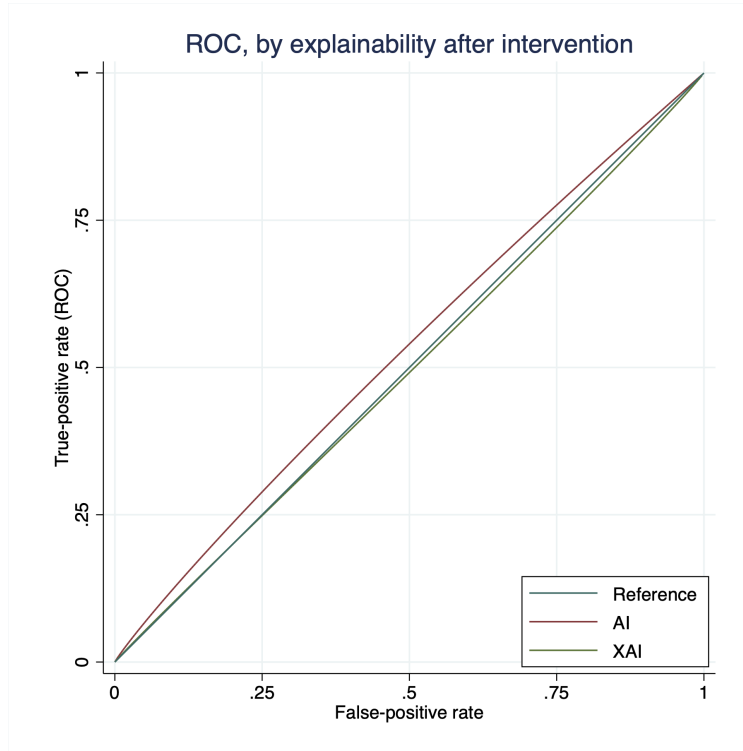
**Figure 17** ROC prior to the treatment intervention.

Notes: We depict ROC plots for our baseline (AI) and treatment (XAI) conditions in the pre-treatment phase, where neither type of participants had access to an AI-based decision aid when making their investment decision. The actual class for the plot is the repayment behavior of an encountered borrower, while the predicted class is participants investment behavior.



**Figure 18** ROC during the treatment intervention.

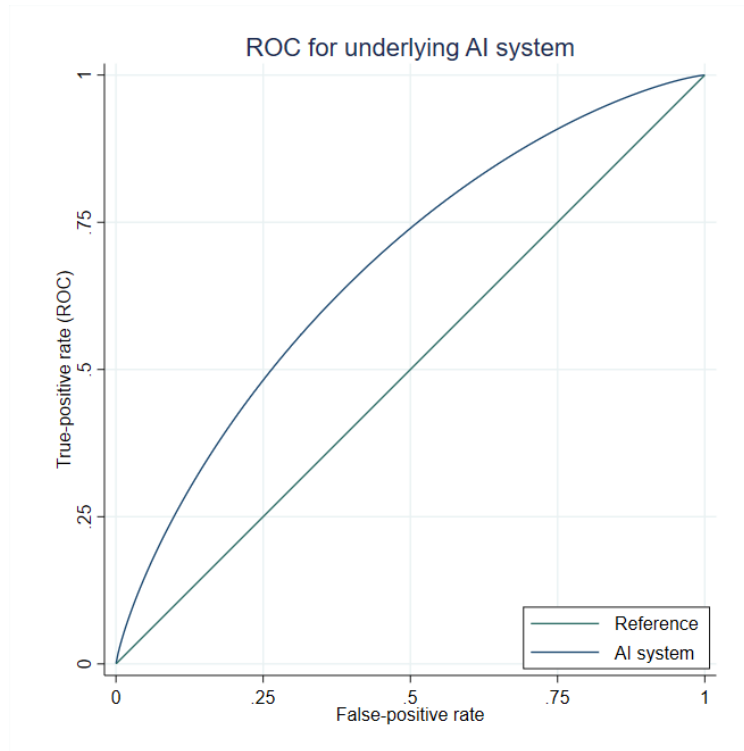
Notes: We depict ROC plots for our baseline (AI) and treatment (XAI) conditions during the treatment phase, where AI participants observed opaque predictions and XAI participants observed explained predictions. The actual class for the plot is the repayment behavior of an encountered borrower, while the predicted class is participants investment behavior.



**Figure 19** ROC after the treatment intervention.

Notes: We depict ROC plots for our baseline (AI) and treatment (XAI) conditions in the in the post-treatment phase, where neither type of participants had access to an AI-based decision aid when making their investment decision. The actual class for the plot is the repayment behavior of an encountered borrower, while the predicted class is participants investment behavior.

The three plots corroborate our finding 1.3 reported in the main text: during and after intervention with the AI system, participants who observed explanations performed significantly worse than those who observed opaque predictions. In the pre-treatment phase, baseline and treatment participants' performance as measured by the ROC-AUC score equaled 0.54 and 0.53 ( $p = 0.18, \chi^2$ -test). During the treatment phase where participants observed predictions, baseline and treatment participants' performance as measured by the ROC-AUC score equaled 0.61 and 0.58 ( $p < 0.01, \chi^2$ -test). Finally, In the post-treatment phase, baseline and treatment participants' performance as measured by the ROC-AUC score equaled 0.55 and 0.52 ( $p < 0.04, \chi^2$ -test). Importantly, as the Figures suggest, baseline participants during and after the treatment intervention outperform their treatment counterparts across the entire range of FPR values.

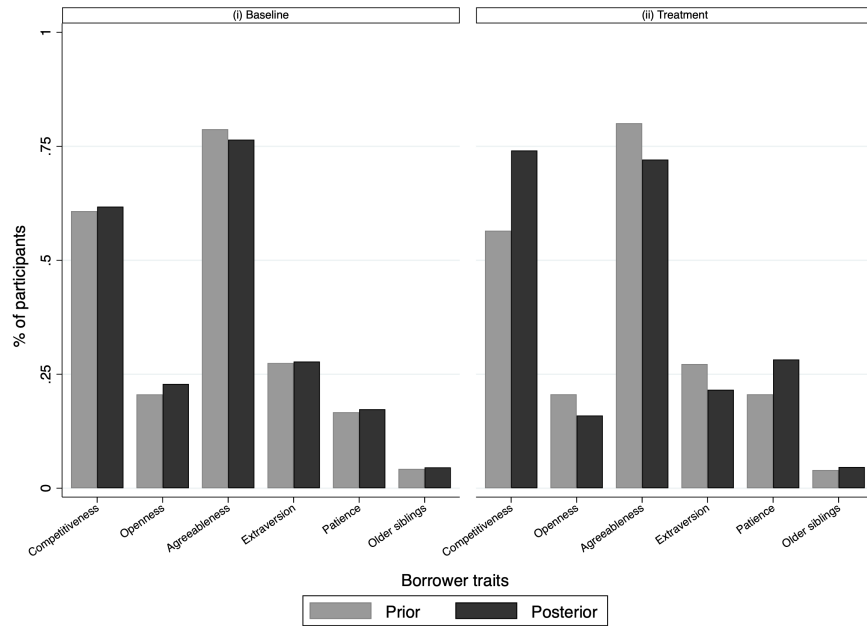


**Figure 20** ROC after the treatment intervention.

Notes: We depict the ROC plot for the underlying AI system's prediction performance in Stage II, i.e., the system's pure performance independent of human participants actual choices. The actual class for the plot is the repayment behavior of an encountered borrower, while the predicted class is the AI systems prediction of the repayment behavior.

Figure 20 depicts the actual (X)AI system's predictive performance. We find that the AI system substantially outperforms human users in the second stage of our experiment: the ROC-AUC score of the AI system in Stage II of study 1 equals 71.6%. This result reveals that the users could have significantly increased their investment performance had they always followed the observed predictions, i.e., machine predictions as such do seem to possess economic value.

**Elicited preferences for observing borrower traits.** We compare participants’ preferences for the borrower traits they want to see before and after they interacted with the AI. For each borrower trait, Figure 21 shows the share of investors who selected it among the three traits to see for their investment decision.<sup>27</sup> Different colored bars represent participant shares before (prior) and after (posterior) participants engaged with the AI. Panel (i) and (ii) portray baseline and treatment results, respectively.



**Figure 21** Preferences over observing borrower traits

Notes: We depict prior and posterior shares of participants who selected a given borrower trait as one of three traits they prefer to see when making the investment decision. Different panels show results for baseline and treatment participants.

Figure 21 corroborates our finding that the provision of explanations not only changes participants’ situational information processing but more fundamentally their conceptions about the relationship between borrower traits and repayment behaviors.

Table 14 depicts regression results that provide additional insights into how the provision of explanations affects participants’ preferences to see borrower traits. We interpret the revealed preference to see a specific borrower trait as the belief about its relevance so that these analyses serve as a robustness check for our result on mental model adjustments (Result 1.2). In all regressions, we use a dummy as a dependent variable that indicates whether participants included a given borrower

<sup>27</sup> 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.

Dep. variable: Including trait in selection	(1)	(2)	(3)	(4)	(5)	(6)
	Competitiveness	Openness	Agreeableness	Extraversion	Patience	Older siblings
XAI	-0.049 (0.041)	0.024 (0.033)	0.011 (0.032)	-0.002 (0.036)	0.045 (0.032)	-0.003 (0.016)
Post	0.003 (0.026)	0.023 (0.023)	-0.020 (0.026)	0.003 (0.028)	0.007 (0.024)	0.003 (0.013)
XAI × Post	0.173*** (0.039)	-0.070** (0.032)	-0.057 (0.037)	-0.060 (0.041)	0.067* (0.039)	0.003 (0.019)
Constant	0.635*** (0.140)	0.101 (0.110)	0.818*** (0.116)	0.256** (0.118)	0.101 (0.100)	0.135* (0.070)
N	1206	1206	1206	1206	1206	1206
p	0.000	0.002	0.000	0.054	0.039	0.048
R <sup>2</sup>	0.044	0.049	0.088	0.034	0.035	0.060

**Table 14** Changes in preferences over observing borrower traits.

Notes: We depict results for OLS regressions. We report robust standard errors in parentheses. In each regression, revealing the preference to see the corresponding borrower trait by selecting it either on place 1, 2, or 3 serves as the dependent (dummy) variable. The independent variables of main interest are a treatment dummy, a dummy indicating the posterior selection decision, and their interaction term. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

trait in their selection of the three traits they want to see. As independent variables, we include a dummy variable controlling for the participation in the XAI treatment, a dummy indicating the posterior selection decision, and their interaction term. Our main interest lies in the interaction term that depicts pure explanation-driven changes in participants' revealed preferences. Note that the reported results are robust to the inclusion of additional participant controls such as gender, education, risk aversion, etc.

We find that observing opaque predictions alone did not entail a significant change in participants' selection of the three borrower traits they want to see (Panel(i) and Table 14). By contrast, Panel (ii) depicts that after observing explanations, participants' preferences to see specific borrower traits changed selectively. Before and after interacting with the XAI, 56.5%, and 74.1% of participants opted to see a borrower's *Competitiveness*. This increase is statistically significant (see column (1) in Table 14). Regarding *Patience* the respective shares equal 20.6% and 28.2%, i.e., the share increases by 7.6% which is statistically significant (see column (5) in Table 14). Considering prior and posterior preferences to see a borrower's *Agreeableness*, we do not find a significant explanation effect (see column (3) in Table 14). Hence, corroborating our results reported in the main text, we find that observing explanations led participants to place more emphasis on a borrower's *Competitiveness* and *Patience* – the traits that explanations depict as highly important to a borrower's repayment behavior – while their preferences over *Agreeableness* – the trait participants initially consider most important and explanations depict as virtually irrelevant – remained unchanged.

**Additional robustness checks.** To ensure that our analyses do not implicitly select against participants who either always or never invest – in the following respectively referred to as types A and B –, we next perform robustness checks. Specifically, we rerun our main regression analyses on subsamples of our data that exclude either or both of these types. Overall, these two types only make up a small minority of our sample. Only 2.5% (3.8%) of our participants always (never) invest, i.e., are of type A (B). In the following, we will report robustness checks for our main results regarding the situational information processing and mental model adjustment process. We always report regression results for subsamples that exclude (i) type A participants, (ii) type B participants, and (iii) type A and B participants. Overall, these analyses reveal that our results reported in the main text are robust to excluding type A, type B, or both. In other words, our results are driven by participants who are neither pure altruists nor players who always play the subgame-perfect strategy of not making an investment. Instead, our results stem from those participants whose behavior suggests that they try to invest with borrowers whom they believe will make a repayment, i.e., individuals who, from a conceptual point of view, should be most inclined to learn to recognize repaying borrowers.

Tables 15, 16, and 17 replicate the analysis reported in Table 9, i.e., situational information processing, for subsamples that exclude type A participants, type B participants, and type A and B participants, respectively. These analyses show that our results on situational information processing are robust to excluding either or both of the aforementioned types, i.e., that selection against certain types of behaviors does not enter into our statistical exercises.

Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage II	Stage I	Stage II	Effect	Effect
Competit.	-0.042*** (0.012)	-0.021*** (0.007)	-0.036*** (0.011)	-0.099*** (0.007)	0.021* (0.012)	-0.09*** (0.017)
Openness	0.025*** (0.009)	0.011** (0.005)	0.03*** (0.008)	0.012* (0.007)	-0.015 (0.01)	-0.001 (0.016)
Agreeabln.	0.086*** (0.010)	0.052*** (0.007)	0.088*** (0.010)	0.008 (0.007)	-0.033*** (0.012)	-0.041** (0.018)
Extrav.	0.029*** (0.009)	0.012* (0.006)	0.035*** (0.009)	0.003 (0.011)	-0.017 (0.011)	0.038 (0.025)
Patience	0.032*** (0.008)	0.007 (0.007)	0.038*** (0.008)	0.03*** (0.009)	-0.026*** (0.01)	0.035** (0.014)
Older sibl.	0.029*** (0.009)	0.000 (0.005)	0.024*** (0.009)	0.027*** (0.009)	-0.029*** (0.011)	0.020 (0.014)
Repayment pred.		0.235*** (0.012)		0.164*** (0.008)		-0.051*** (0.016)
N	2910	5820	2930	5860	9180	17520
p	0.000	0.000	0.000	0.000	0.000	0.000
R <sup>2</sup>	0.385	0.453	0.446	0.410	0.386	0.430

**Table 15** Change in information weighting across Stages I and II – robustness check.

Notes: We depict results for OLS regressions with fixed effects for a subsample that excludes participants who always invest their 10 MU. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and II serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and II, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage II	Stage I	Stage II	Effect	Effect
Competit.	-0.04*** (0.011)	-0.021*** (0.007)	-0.036*** (0.011)	-0.098*** (0.007)	0.018 (0.012)	-0.085*** (0.019)
Openness	0.024*** (0.008)	0.01* (0.005)	0.029*** (0.009)	0.013** (0.007)	-0.014 (0.01)	-0.000 (0.015)
Agreeabln.	0.083*** (0.010)	0.051*** (0.007)	0.087*** (0.010)	0.009 (0.007)	-0.032*** (0.012)	-0.04** (0.018)
Extrav.	0.028*** (0.009)	0.011* (0.006)	0.035*** (0.009)	0.000 (0.011)	-0.017 (0.011)	0.036 (0.024)
Patience	0.032*** (0.008)	0.005 (0.007)	0.039*** (0.008)	0.028*** (0.009)	-0.027*** (0.01)	0.035** (0.014)
Older sibl.	0.028*** (0.009)	0.001 (0.005)	0.024*** (0.009)	0.026*** (0.009)	-0.028*** (0.01)	0.018 (0.015)
Repayment pred.		0.229*** (0.012)		0.177*** (0.011)		-0.05*** (0.016)
N	2990	5980	2930	5860	8970	17760
p	0.000	0.000	0.000	0.000	0.000	0.000
R <sup>2</sup>	0.361	0.444	0.417	0.396	0.415	0.414

**Table 16** Change in information weighting across Stages I and II – robustness check.

Notes: We depict results for OLS regressions with fixed effects for a subsample that excludes participants who never invest their 10 MU. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and II serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and II, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .



Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage II	Stage I	Stage II	Effect	Effect
Competit.	-0.043*** (0.012)	-0.022*** (0.007)	-0.037*** (0.011)	-0.101*** (0.007)	0.02 (0.013)	-0.09*** (0.02)
Openness	0.026*** (0.009)	0.01* (0.006)	0.029*** (0.009)	0.013* (0.007)	-0.015 (0.01)	-0.001 (0.016)
Agreeabln.	0.087*** (0.010)	0.055*** (0.008)	0.091*** (0.011)	0.008 (0.007)	-0.033*** (0.012)	-0.043** (0.019)
Extrav.	0.03*** (0.009)	0.011* (0.006)	0.036*** (0.009)	0.001 (0.011)	-0.019* (0.011)	0.036 (0.025)
Patience	0.034*** (0.008)	0.007 (0.007)	0.04*** (0.008)	0.029*** (0.009)	-0.027*** (0.01)	0.034** (0.015)
Older sibl.	0.03*** (0.009)	0.001 (0.005)	0.025*** (0.009)	0.028*** (0.009)	-0.029*** (0.011)	0.019 (0.015)
Repayment pred.		0.24*** (0.012)		0.182*** (0.011)		-0.057*** (0.017)
N	2840	5680	2850	5700	8520	17070
p	0.000	0.000	0.000	0.000	0.000	0.000
R <sup>2</sup>	0.346	0.437	0.412	0.39	0.408	0.408

**Table 17** Change in information weighting across Stages I and II – robustness check.

Notes: We depict results for OLS regressions with fixed effects for a subsample that excludes participants who always or never invest their 10 MU. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and II serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and II, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Tables 18, 19, and 20 replicate the analysis reported in Table 12, i.e., mental model adjustments, for subsamples that exclude type A participants, type B participants, and type A and B participants, respectively. These analyses show that our results on mental model adjustments are robust to excluding either or both of the aforementioned types, i.e., that selection against certain types of behaviors does not enter into our statistical exercises.

Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage III	Stage I	Stage III	Effect	Effect
Competit.	-0.042*** (0.012)	-0.047*** (0.012)	-0.036*** (0.011)	-0.089*** (0.013)	-0.005 (0.014)	-0.048** (0.02)
Openness	0.025*** (0.009)	0.027*** (0.01)	0.03*** (0.009)	-0.001 (0.009)	0.001 (0.011)	-0.032** (0.016)
Agreeabln.	0.086*** (0.010)	0.102*** (0.011)	0.088*** (0.010)	0.081*** (0.010)	0.017 (0.012)	-0.024 (0.017)
Extrav.	0.029*** (0.009)	0.048*** (0.010)	0.035*** (0.009)	0.025** (0.010)	0.019* (0.011)	-0.029* (0.017)
Patience	0.032*** (0.008)	0.017* (0.009)	0.038*** (0.008)	0.061*** (0.010)	-0.016 (0.010)	0.038** (0.015)
Older sibl.	0.029*** (0.009)	0.035*** (0.01)	0.024*** (0.009)	0.018** (0.009)	0.006 (0.011)	-0.011 (0.016)
N	2910	2910	2930	2930	5820	11680
p	0.000	0.000	0.000	0.000	0.000	0.000
R <sup>2</sup>	0.37	0.373	0.441	0.385	0.371	0.393

**Table 18** Change in information weighting across Stages I and III – robustness check.

Notes: We depict results for OLS regressions with fixed effects for a subsample that excludes participants who always invest their 10 MU. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and III serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and III, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage III	Stage I	Stage III	Effect	Effect
Competit.	-0.04*** (0.012)	-0.044*** (0.012)	-0.036*** (0.011)	-0.089*** (0.013)	-0.005 (0.013)	-0.049** (0.019)
Openness	0.024*** (0.008)	0.026*** (0.01)	0.029*** (0.009)	-0.002 (0.009)	0.001 (0.011)	-0.032** (0.016)
Agreeabln.	0.083*** (0.010)	0.099*** (0.011)	0.087*** (0.010)	0.08*** (0.010)	0.016 (0.012)	-0.023 (0.016)
Extrav.	0.028*** (0.009)	0.046*** (0.010)	0.035*** (0.009)	0.026** (0.010)	0.018 (0.011)	-0.027 (0.017)
Patience	0.032*** (0.008)	0.017* (0.009)	0.039*** (0.008)	0.061*** (0.010)	-0.015 (0.010)	0.037** (0.016)
Older sibl.	0.028*** (0.009)	0.034*** (0.009)	0.024*** (0.009)	0.019** (0.009)	0.006 (0.011)	-0.011 (0.015)
N	2910	2910	2930	2930	5820	11680
p	0.000	0.000	0.000	0.000	0.000	0.000
R <sup>2</sup>	0.37	0.373	0.441	0.385	0.371	0.393

**Table 19** Change in information weighting across Stages I and III – robustness check.

Notes: We depict results for OLS regressions with fixed effects for a subsample that excludes participants who never invest their 10 MU. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and III serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and III, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Dep. variable:	Baseline		Treatment		(5)	(6)
	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage III	Stage I	Stage III	Effect	Effect
Competit.	-0.043*** (0.012)	-0.048*** (0.012)	-0.037*** (0.011)	-0.092*** (0.013)	-0.005 (0.014)	-0.05** (0.02)
Openness	0.026*** (0.009)	0.027*** (0.01)	0.029*** (0.009)	-0.002 (0.009)	0.001 (0.012)	-0.033** (0.016)
Agreeabln.	0.087*** (0.010)	0.104*** (0.011)	0.091*** (0.011)	0.084*** (0.010)	0.017 (0.012)	-0.024 (0.017)
Extrav.	0.03*** (0.01)	0.049*** (0.010)	0.036*** (0.009)	0.027*** (0.010)	0.019* (0.011)	-0.028 (0.017)
Patience	0.034*** (0.009)	0.018* (0.01)	0.04*** (0.008)	0.063*** (0.010)	-0.016 (0.011)	0.038** (0.016)
Older sibl.	0.03*** (0.009)	0.036*** (0.009)	0.025*** (0.009)	0.019** (0.009)	0.006 (0.011)	-0.012 (0.016)
N	2840	2840	2850	2850	5680	11380
p	0.000	0.000	0.000	0.000	0.000	0.000
R <sup>2</sup>	0.346	0.35	0.412	0.363	0.348	0.369

**Table 20** Change in information weighting across Stages I and III – robustness check.

Notes: We depict results for OLS regressions with fixed effects for a subsample that excludes participants who always or never invest their 10 MU. We report robust standard errors in parentheses. Participants' investment decisions in Stages I and III serve as the dependent variable. As independent variables, we include all borrower traits, LIME values that do not create multicollinearity, and the borrower's actual type. Columns (1) and (2) show estimates for baseline participants' decisions in Stages I and III, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows estimated differences between (1) and (2), measuring weight changes driven by the provision of opaque predictions. Finally, column (6) shows DiD estimates, revealing explanation-driven weight changes. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Tables 21, 22, and 23 replicate the analysis reported in Table 13, i.e., the investment performance of participants, for subsamples that exclude type A participants, type B participants, and type A and B participants, respectively. These analyses show that our results on participants' investment performance are robust to excluding either or both of the aforementioned types, i.e., that selection against certain types of behaviors does not enter into our statistical exercises.

	Accuracy			Recall		
	(1) Stage I	(2) Stage II	(3) Stage III	(4) Stage I	(5) Stage II	(6) Stage III
XAI ( $\alpha_1$ )	0.012 (0.020)	-0.012 (0.016)	-0.027 (0.020)	0.041 (0.027)	-0.01 (0.020)	-0.009 (0.026)
Very high Competit. ( $\alpha_2$ )	-0.077*** (0.027)	-0.043** (0.017)	-0.107*** (0.028)	0.008 (0.035)	-0.012 (0.021)	0.02 (0.037)
XAI $\times$ Very high Competit. ( $\alpha_3$ )	-0.005 (0.026)	-0.108*** (0.019)	-0.025 (0.027)	-0.019 (0.032)	-0.147*** (0.024)	-0.086** (0.034)
F-test: $\alpha_1 + \alpha_3 = 0$	$p = 0.761$	$p < 0.01$	$p < 0.03$	$p = 0.482$	$p < 0.01$	$p < 0.01$
N	5840	11680	5840	4523	9386	4523
p	0.000	0.000	0.000	0.000	0.000	0.000
$R^2$	0.023	0.14	0.024	0.05	0.237	0.07

**Table 21 Treatment differences for different performance measures – robustness check.**

Notes: We depict results for OLS regressions for a subsample that excludes participants who always invest their 10 MU. We report robust standard errors in parentheses. In columns (1) to (3), we use a dummy as the dependent variable that indicates whether a participant made the payoff maximizing investment decision – Accuracy. In columns (4) to (6), we use a dummy as the dependent variable that indicates whether a participant correctly invested with a repaying borrower – Recall. The independent variables of main interest are a treatment dummy, a dummy indicating that the borrower is most competitive, and their interaction term. We additionally control for borrowers' other traits, and, if appropriate, for the observed prediction and LIME values. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

	Accuracy			Recall		
	(1) Stage I	(2) Stage II	(3) Stage III	(4) Stage I	(5) Stage II	(6) Stage III
XAI ( $\alpha_1$ )	0.006 (0.019)	-0.011 (0.024)	-0.032* (0.019)	0.035 (0.025)	-0.012 (0.018)	-0.015 (0.024)
Very high Competit. ( $\alpha_2$ )	-0.102*** (0.027)	-0.034** (0.017)	-0.131*** (0.028)	0.007 (0.035)	-0.01 (0.022)	0.005 (0.037)
XAI $\times$ Very high Competit. ( $\alpha_3$ )	-0.002 (0.026)	-0.113*** (0.019)	-0.023 (0.026)	-0.021 (0.031)	-0.155*** (0.024)	-0.088*** (0.033)
F-test: $\alpha_1 + \alpha_3 = 0$	$p = 0.858$	$p < 0.01$	$p < 0.02$	$p = 0.652$	$p < 0.01$	$p < 0.01$
N	5920	11840	5920	4576	9510	4576
p	0.000	0.000	0.000	0.000	0.000	0.000
$R^2$	0.027	0.147	0.025	0.05	0.243	0.071

**Table 22 Treatment differences for different performance measures – robustness check.**

Notes: We depict results for OLS regressions for a subsample that excludes participants who never invest their 10 MU. We report robust standard errors in parentheses. In columns (1) to (3), we use a dummy as the dependent variable that indicates whether a participant made the payoff maximizing investment decision – Accuracy. In columns (4) to (6), we use a dummy as the dependent variable that indicates whether a participant correctly invested with a repaying borrower – Recall. The independent variables of main interest are a treatment dummy, a dummy indicating that the borrower is most competitive, and their interaction term. We additionally control for borrowers' other traits, and, if appropriate, for the observed prediction and LIME values. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

	Accuracy			Recall		
	(1) Stage I	(2) Stage II	(3) Stage III	(4) Stage I	(5) Stage II	(6) Stage III
XAI ( $\alpha_1$ )	0.012 (0.019)	-0.008 (0.015)	-0.028 (0.019)	0.043* (0.025)	-0.006 (0.018)	-0.009 (0.025)
Very high Competit. ( $\alpha_2$ )	-0.097*** (0.028)	-0.04** (0.018)	-0.127*** (0.028)	-0.01 (0.035)	-0.004 (0.023)	0.003 (0.038)
XAI $\times$ Very high Competit. ( $\alpha_3$ )	-0.003 (0.026)	-0.112*** (0.019)	-0.025 (0.027)	-0.016 (0.032)	-0.153*** (0.024)	-0.085*** (0.034)
F-test: $\alpha_1 + \alpha_3 = 0$	$p = 0.706$	$p < 0.01$	$p < 0.03$	$p = 0.389$	$p < 0.01$	$p < 0.01$
N	5690	11380	5690	4402	9144	4402
p	0.000	0.000	0.000	0.000	0.000	0.000
$R^2$	0.027	0.152	0.026	0.054	0.255	0.075

**Table 23 Treatment differences for different performance measures – robustness check.**

Notes: We depict results for OLS regressions for a subsample that excludes participants who always or never invest their 10 MU. We report robust standard errors in parentheses. In columns (1) to (3), we use a dummy as the dependent variable that indicates whether a participant made the payoff maximizing investment decision – Accuracy. In columns (4) to (6), we use a dummy as the dependent variable that indicates whether a participant correctly invested with a repaying borrower – Recall. The independent variables of main interest are a treatment dummy, a dummy indicating that the borrower is most competitive, and their interaction term. We additionally control for borrowers' other traits, and, if appropriate, for the observed prediction and LIME values. Estimates are standardized. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

## Study 2: Analyses

**Prior beliefs and absolute belief adjustments.** Table 24 shows results for different regression models. In column (1) the dependent variable is participants’ beliefs about the contribution of apartment attributes to apartments’ listing prices in Stage I. In column (2) the dependent variable is the absolute difference between these beliefs elicited in Stages I and III, i.e., before and after the treatment intervention. In both columns, the independent variables of main interest are dummies indicating whether participants observed predictions in Stage II (Prediction) and on top of predictions SHAP explanations (SHAP). We additionally included participants’ controls, and apartment fixed effects. We report robust standard errors in parentheses.

	(1) Prior belief	(2) Abs. belief adjustment
Prediction ( $\alpha_1$ )	51.617 (33.397)	-3.713 (51.199)
Expl. ( $\alpha_2$ )	20.513 (32.298)	135.410*** (40.726)
F-test: $\alpha_1 + \alpha_2 = 0$	$p = 0.33$	$p < 0.01$
N	1836	1836
p	0.009	0.000
$R^2$	0.115	0.04

**Table 24** Differences in prior beliefs and absolute belief adjustments.

Notes: We depict results for OLS regressions with apartment fixed effects. We report robust standard errors in parentheses. In column (1) and (3), we respectively use participants’ prior belief about the marginal contribution of apartment attributes to the listing price in euros, and their absolute change in a belief as dependent variables. As independent variables, we include a dummy indicating that participants observed a prediction in Stage II (Prediction), and a dummy indicating that they observed SHAP explanations in Stage II (Expl.). We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Regression results reveal that there do not exist significant treatment differences in prior beliefs (see column (1)). However, estimates in column (2) show that observing explanations on top of predictions significantly increases the absolute adjustment of beliefs by about 135€ on average. These observations suggest that explanations evoke belief adjustments for real-estate experts.

**Robustness of confirmation bias measures.** Table 25 depicts regression results that serve as a robustness check regarding the presence of confirmation bias in the mental model adjustment processes. We repeat the regression exercises from table 3 in the main text. We regress participants’ posterior beliefs on their prior beliefs, the observed average SHAP values, a dummy indicating that average SHAP values confirm prior beliefs, and their interaction effects. We report robust standard errors in parentheses. Importantly, and in contrast to the main text analyses, we define that explanations confirm prior beliefs in a more restrictive way: explanations confirm priors if the absolute distance between the prior and the observed average SHAP value is smaller than the absolute distance between the prior and 0€ and between the prior and the closest extreme, i.e., +/- 2500€.

Dep. variable:	(1)	(2)	(3)
Posterior belief	Overall	Low confidence beliefs	High confidence beliefs
Prior belief	0.496*** (0.061)	0.463*** (0.070)	0.748*** (0.105)
Avg. SHAP	0.397*** (0.028)	0.424*** (0.037)	0.249*** (0.047)
Confirm	-21.026 (30.661)	-51.430 (45.167)	131.448* (77.220)
Prior belief × Confirm	0.117 (0.093)	0.250** (0.112)	-0.342* (0.190)
Avg. SHAP × Confirm	-0.123 (0.087)	-0.324** (0.139)	0.231** (0.110)
N	708	481	222
p	0.000	0.000	0.000
R <sup>2</sup>	0.743	0.728	0.840

**Table 25 Confirmation bias and posterior belief formation – Robustness check**

Notes: We depict results from OLS regression models with individual and apartment fixed effects. We report robust standard errors reported in parentheses. The dependent variable equals XAI participants’ posterior belief about the marginal contribution of apartment attributes to the listing price in euros. The main independent variables of interest are participants’ prior beliefs, the average SHAP values for apartment attributes in Stage II, a dummy indicating that observed SHAP values in Stage II confirmed participants’ priors – explanations confirm priors if the absolute distance between the prior and the observed average SHAP value is smaller than the absolute distance between the prior and 0€ and between the prior and the closest extreme, i.e., +/- 2500€ – and interaction terms. We further control for the overall posterior listing price participants entered for the apartment and the average prediction they observed in Stage II. Column (1) presents results for all decisions. Columns (2) and (3) respectively depict results for the shares of decisions where XAI participants report low and high confidence in their prior. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Even with this more restrictive definition of confirming explanations, we continue to find evidence for the presence of confirmation bias. Namely, for high confidence compared to low confidence beliefs, experts are generally less inclined to adjust beliefs in the direction of the observed explanation. However, when the SHAP values confirm their priors, they are significantly more inclined to change beliefs according to explanations. By contrast, for prior beliefs where participants report low confidence, we find that they adjust their beliefs more strongly in the direction of the explanation, in case the explanation was contradicting their prior. Hence, as the literature suggests (see, e.g., Knobloch-Westerwick and Meng 2009), the confirmatory adjustment of beliefs is considerably more pronounced for beliefs in which experts have high confidence.

**Spillover effects.** Table 26 shows results for regression analyses on participants listing price estimations for the apartment in Chemnitz they observe at the end of the study. The dependent variable is the entered listing price estimate.

Dep. variable: Price estimate Chemnitz	(1) Overall	(2) Below Median belief adjustment	(3) Above Median belief adjustment
Balcony $\times$ AI	-298.992 (801.443)	-233.176 (884.999)	-2073.208 (1736.720)
Balcony $\times$ Expl.	-79.185 (746.115)	-687.224 (1567.682)	283.148 (1339.158)
Low green $\times$ AI	894.361 (993.437)	483.990 (1086.211)	2575.297 (2197.200)
High green $\times$ AI	-390.458 (1064.479)	717.449 (1019.652)	-2551.901 (2056.778)
Low green $\times$ Expl.	-1902.323** (877.642)	495.868 (2091.666)	-3459.922** (1632.984)
High green $\times$ Expl.	1742.126* (911.520)	84.433 (1736.347)	3906.959*** (1312.532)
N	153	72	81
p	0.000	0.000	0.000
$R^2$	0.289	0.471	0.527

**Table 26** Listing price estimation for apartments in Chemnitz.

Notes: We depict results from OLS regression models with individual and apartment fixed effects. We report robust standard errors reported in parentheses. The dependent variable equals the listing price estimate for a Chemnitz apartment in euros. The main independent variables of interest are dummies indicating that the participants observed predictions in Stage II (AI), that the participants observed explanations in Stage II (Expl.), that the evaluated apartment has a balcony (Balcony), that the evaluated apartment is in a district where the share of green voters is low (Low green), that the evaluated apartment is in a district where the share of green voters is high (High green), and their interaction effects. As additional controls, we include participants' age, experience in the real estate industry, experience with estimating listing prices, general overconfidence, contextualized overconfidence for the task, risk aversion, familiarity with AI decision support, gender, and education level. Column (1) depicts results across all participants. Columns (2) and (3) respectively show results for regression analyses performed on the subsample of participants whose belief adjustment across Stages I and III for the "Green Voter" attribute lies below and above the median. We denote significance levels by \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

In all columns, the independent variables of main interest are dummies indicating that the participants observed predictions in Stage II (AI), that the participants observed explanations in Stage II (Expl.), that the evaluated apartment has a balcony (Balcony), that the evaluated apartment is in a district where the share of green voters is low (Low green), that the evaluated apartment is in a district where the share of green voters is high (High green), and their interaction effects. As additional controls, we include participants' reported socio-demographics. Columns (2) and (3) respectively show results for regression analyses performed on the subsample of participants whose belief adjustment across Stages I and III for the "Green Voter" attribute lies below and above the median.

Regression results reveal significant explanation effects regarding the "Green Voter" attribute. The estimates for *Low green*  $\times$  *XAI* and *High green*  $\times$  *XAI* are both statistically significant in column (1). Hence, observing explanations for Cologne and Frankfurt in Stage II led experts to change their strategy of estimating listing prices for an apartment in Chemnitz. Results in columns (2) and (3) further reveal that these effects are driven by experts who strongly adjusted their beliefs for this attribute across Stages I and III.



## 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