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 1 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.

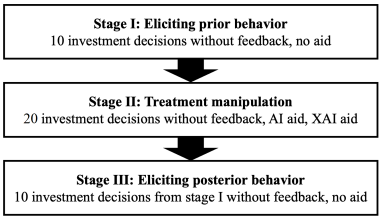


Figure 1 Sequence of the experiment

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 ten deliberately abstract borrower characteristics and decides whether or not to invest her 10 MU with the borrower. Notably, we chose characteristics that previous empirical analyses revealed to be correlated with borrowers' repayment behaviors so that a machine learning model can leverage them to make sufficiently accurate predictions using the same information that participants observe, too. If the investor keeps her endowment, both the investor and borrower receive a payoff of 10 MU. If she invests her endowment, the borrower receives 20 MU and has to decide whether or not to repay the investor by giving up 10 MU. In case of repayment, the investor receives 20 MU so that the initial investment pays off; otherwise the investor ends up with 0 MU while the borrower earns 30 MU (see Figure 2).

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

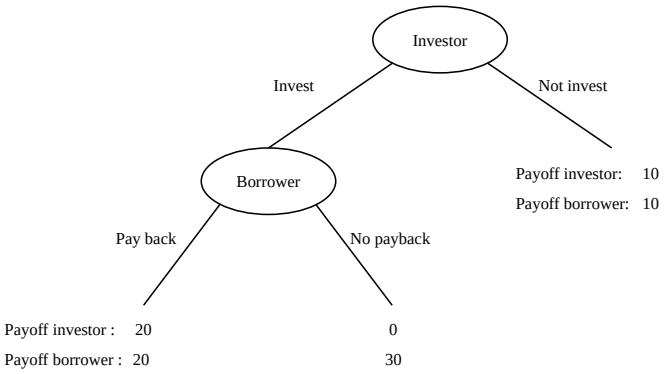


Figure 2 Investment game structure

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

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).

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 3 and Figure 4, 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

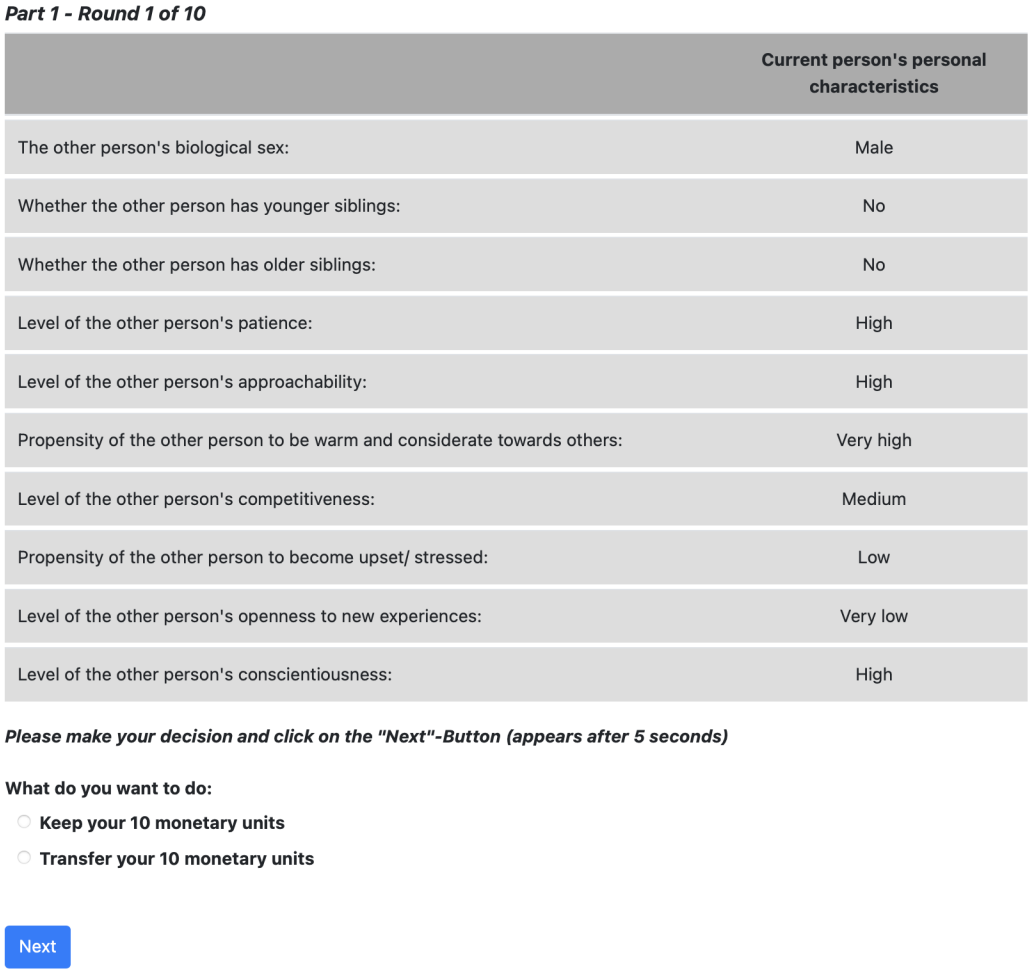


Figure 3 Interface of Study 1 in Stage I and III.

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.

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.

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).[1] 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 1 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

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

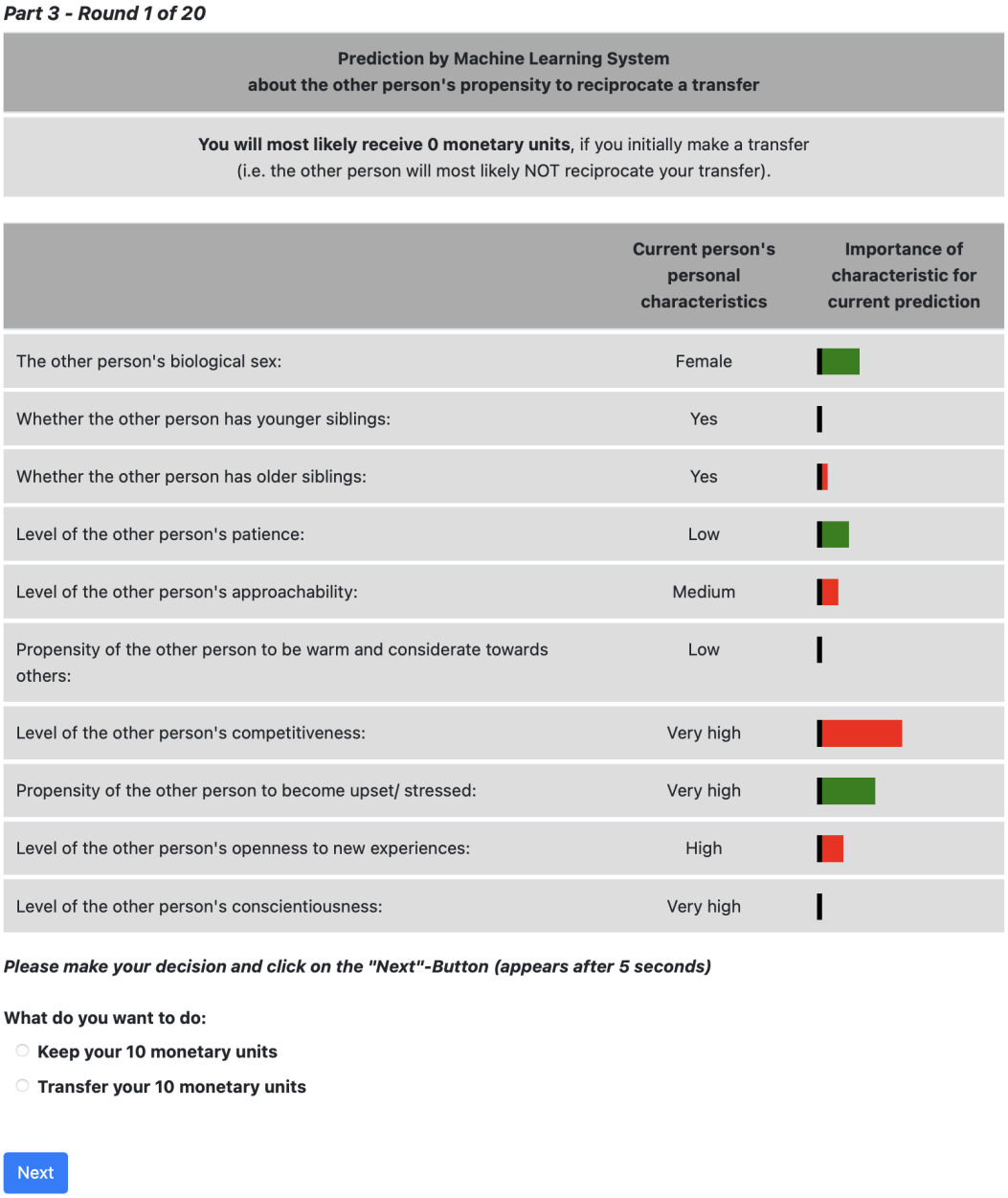


Figure 4 Interface of Study 1 in Stage II, XAI treatment

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.

interacts with 32 borrowers (the same 10 in Stages I and III, 20 in Stage II, and 2 for elciting prior 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 1 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 1 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).[2]

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

[2] 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.[3]

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

[3] 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.

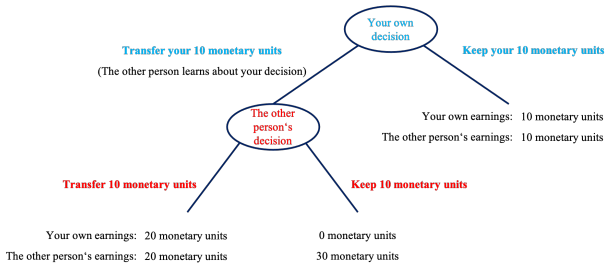
Completion After participants have made all investment decisions, the experiment ends with a questionnaire containing items on participants' socio-demographics and social preferences. Participants' answers serve as controls for some of our regression analyses. At the end of the experiment, we inform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

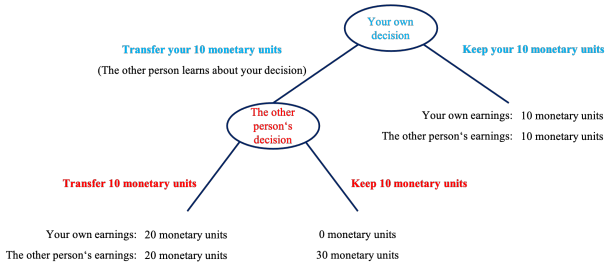
- If the other person transfers 10 monetary units back to you, we double this amount so that you receive 20 monetary units. In this case, your personal and the other person's earnings in this round both equal 20 monetary units.
- If the other person does <u>not</u> transfer 10 monetary units back to 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 <u>not</u> 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:

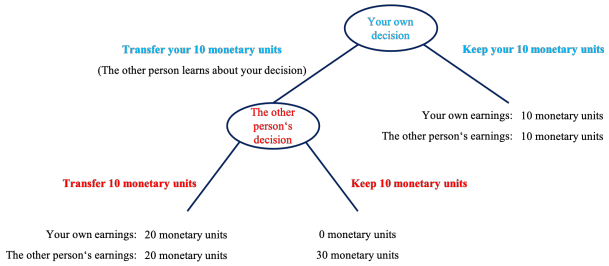
- **First choice:** The other person's characteristic you select as the first choice will be shown to you with a probability of 100%.
- **Second choice:** The other person's characteristic you select as the second choice will be shown to you with a probability of 90%. With a probability of 10% you will 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 <u>not</u> 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.

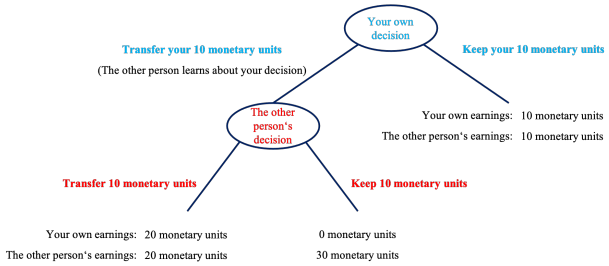
Now that you have finished all 20 rounds, we ask you to guess how good the Machine Learning System's predictions are. Overall you have to make three distinct guesses. For every guess that is not off by more than +/- 10 percentage points from the actual value, you receive 5 monetary units.

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 <u>not</u> 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:

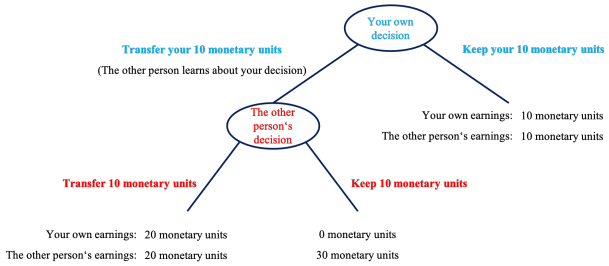
- **First choice:** The other person's characteristic you select as the first choice will be shown to you with a probability of 100%.
- **Second choice:** The other person's characteristic you select as the second choice will be shown to you with a probability of 90%. With a probability of 10% you will 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 <u>not</u> 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.

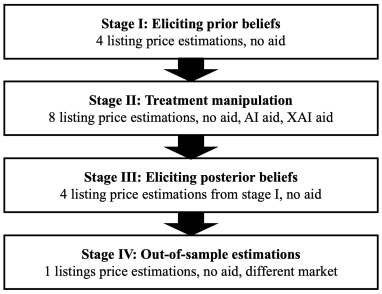


Figure 5 Sequence of Study 2

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

The experimental protocol comprises 4 stages (see Figure 5 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 2 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 2 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.

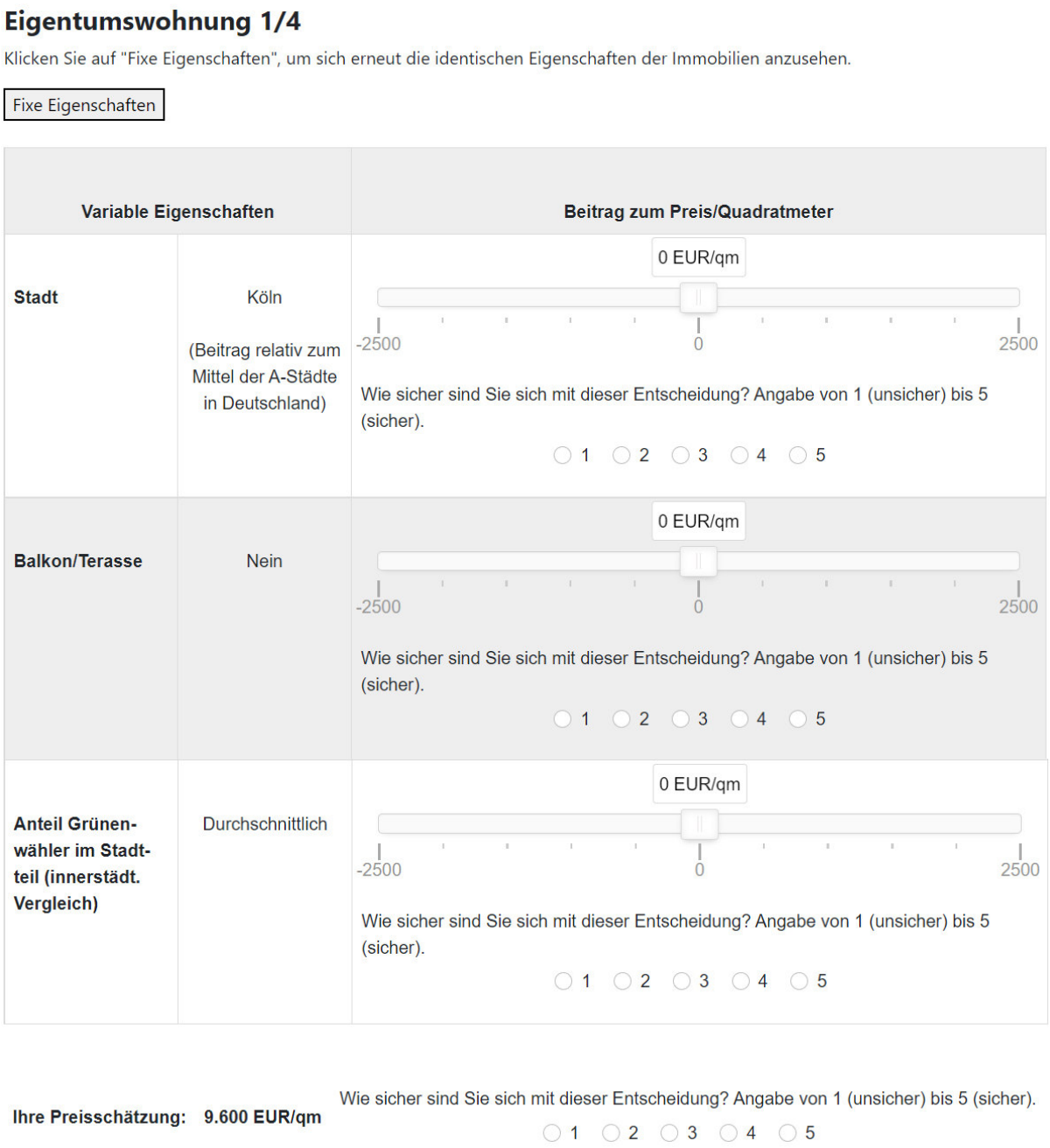
The listings participants encounter are real apartments that we scraped from a large German real-estate platform (www.immonet.de) over a period of 3 weeks in February 2022. The entire data set we collected comprises 5090 distinct observations. Excluding the observations that participants encounter in 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 6 (original) and Figure 7 (English translation).

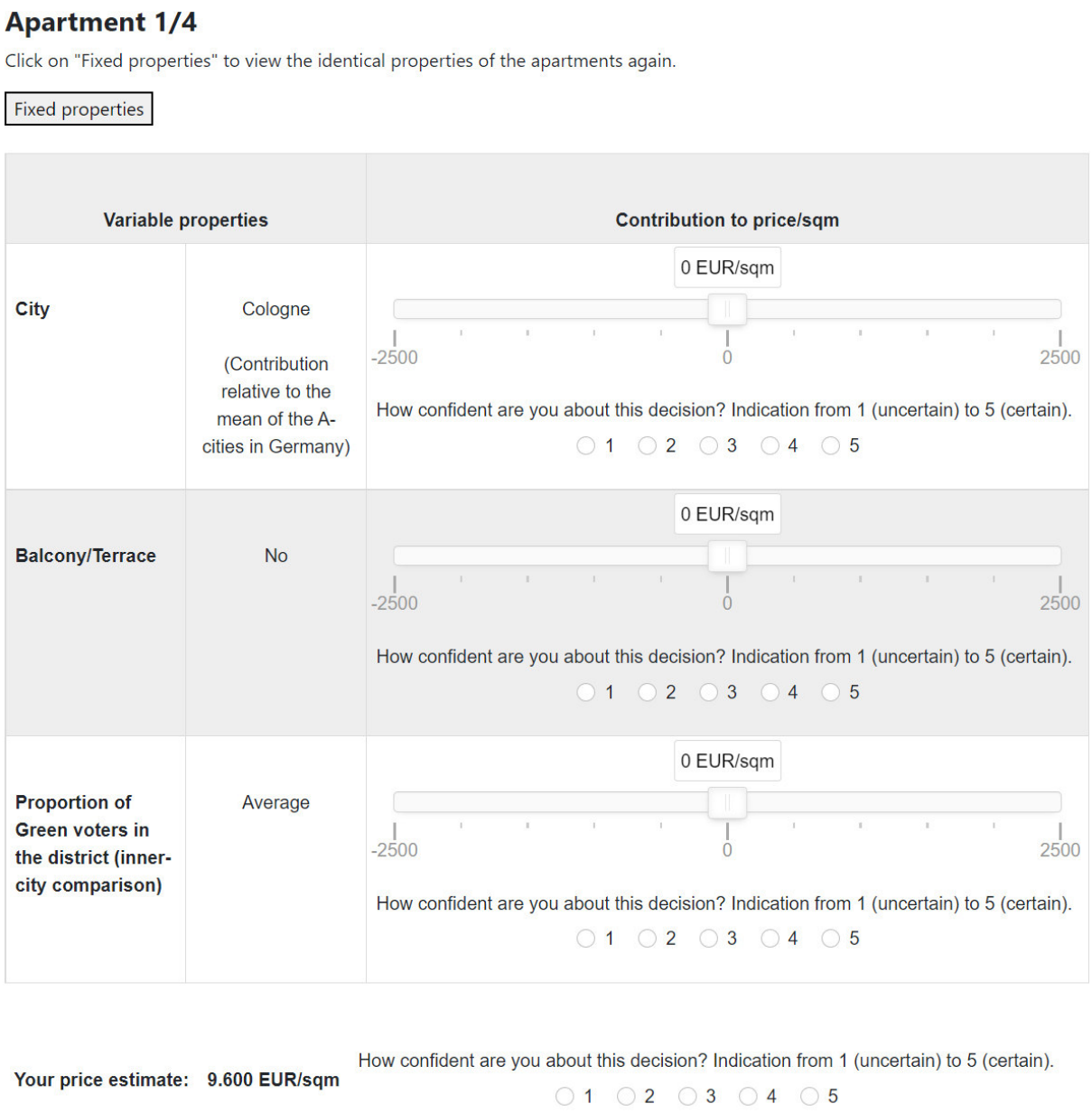
Stage II. In Stage II, participants have to predict the listing price for 8 listings. In contrast to Stage I, participants do not have to enter the contribution for the three variable apartment attributes. Instead, they only predict the overall listing price. We again ask participants to state their 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



(a) Original

Figure 6 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.



(a) English translation

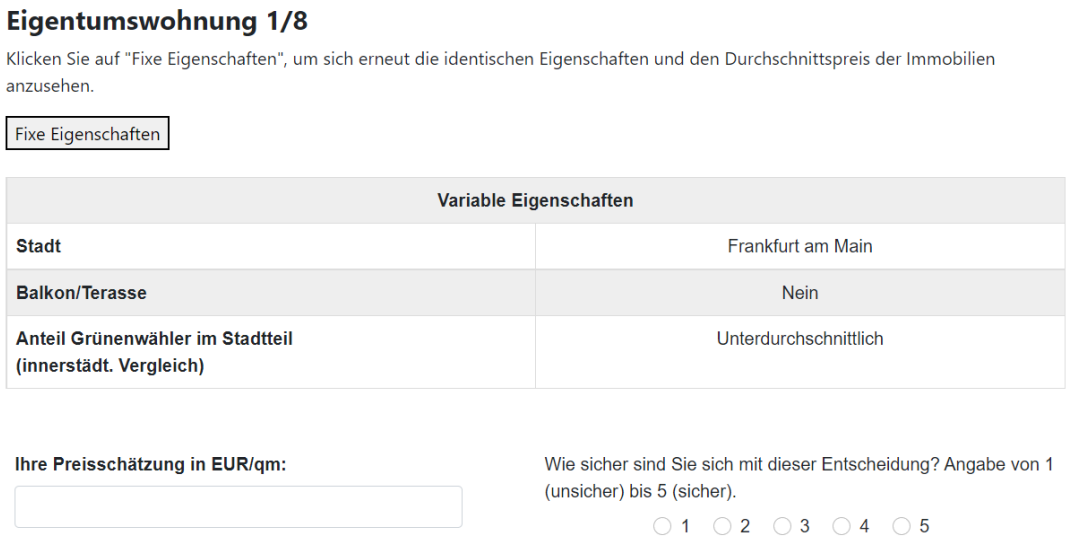
Figure 7 Stage I and 3: Belief elicitation (English translation).

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.

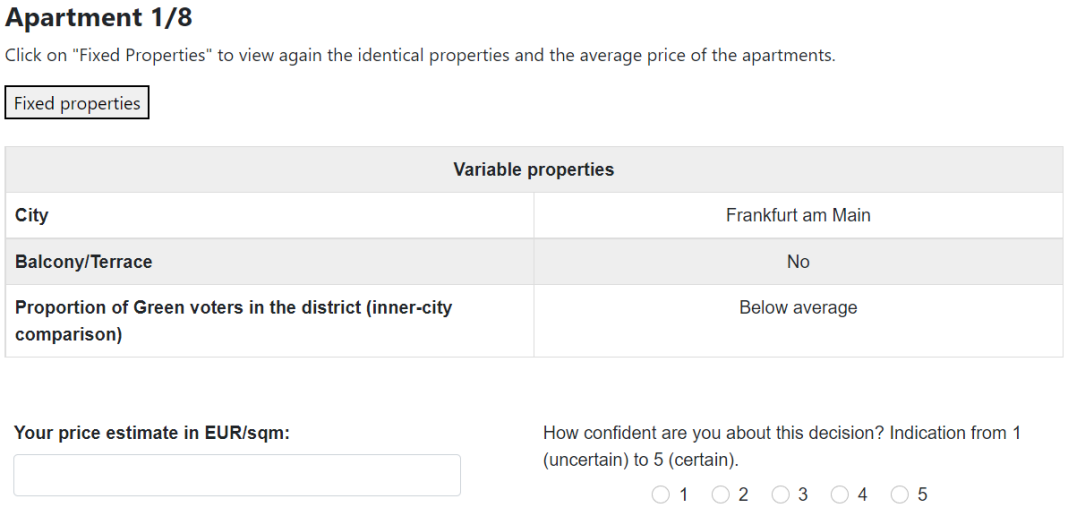
the AI-system's overall price prediction, participants also receive local SHAP explanations. More specifically, for every single listing they encounter, participants in the XAI condition observe the instance's idiosyncratic SHAP values for the three variable apartment characteristics. We depict the local 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 8 (NoAid), in Figure 9 (AI), and in Figure 10 (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 11. After Stage IV, the study concludes with a brief questionnaire on participants' socio-demographics including their age, gender, and working experience.



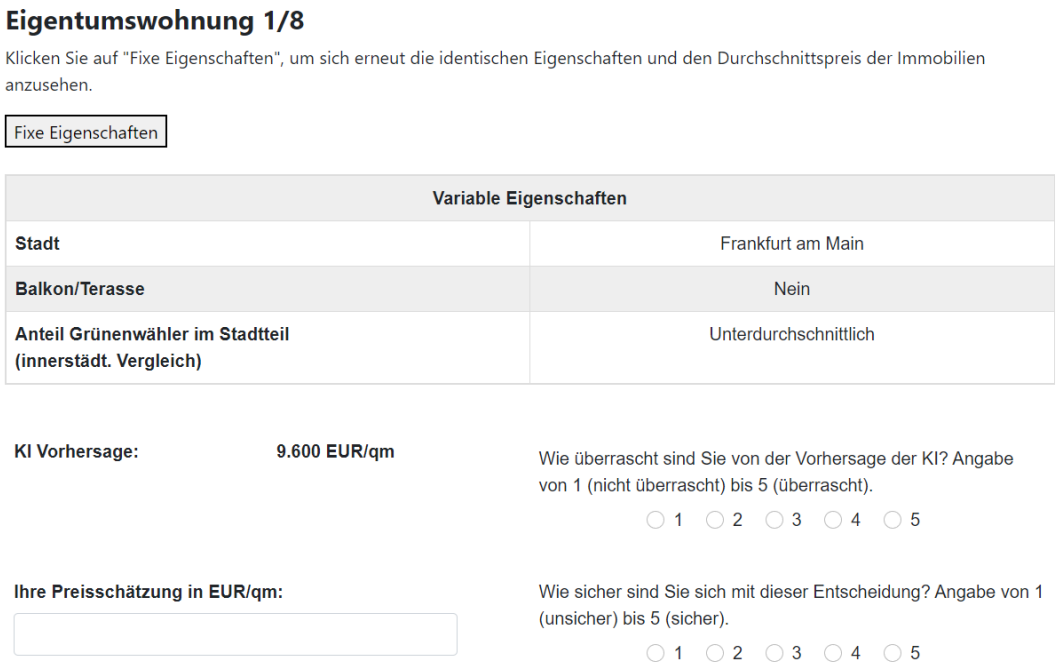
(a) Original



(b) English translation

Figure 8 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).

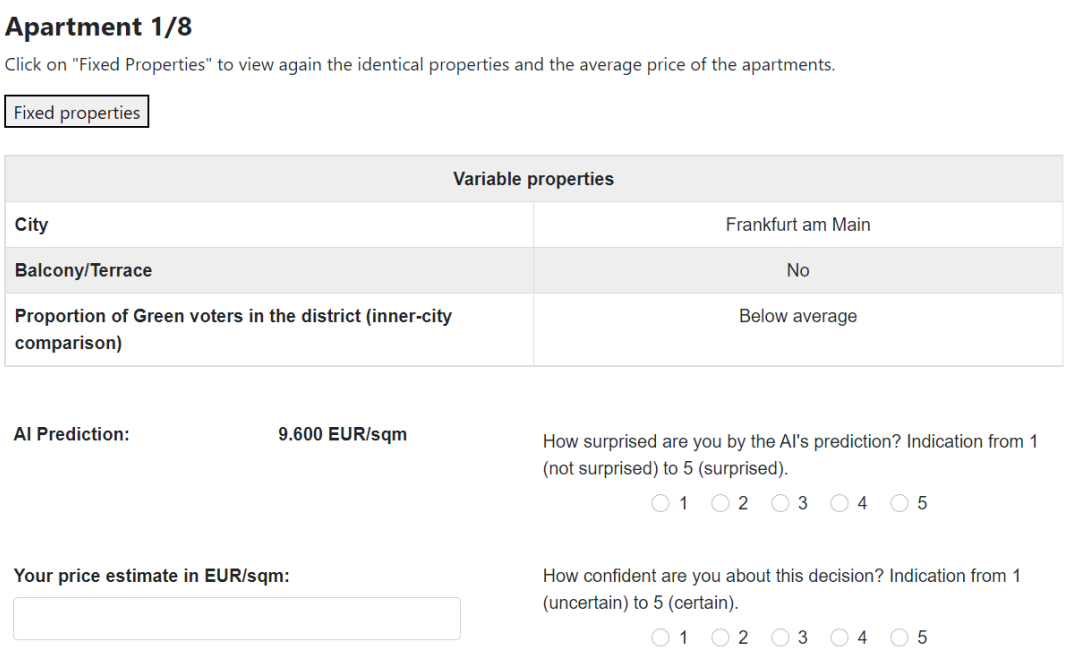
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

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:

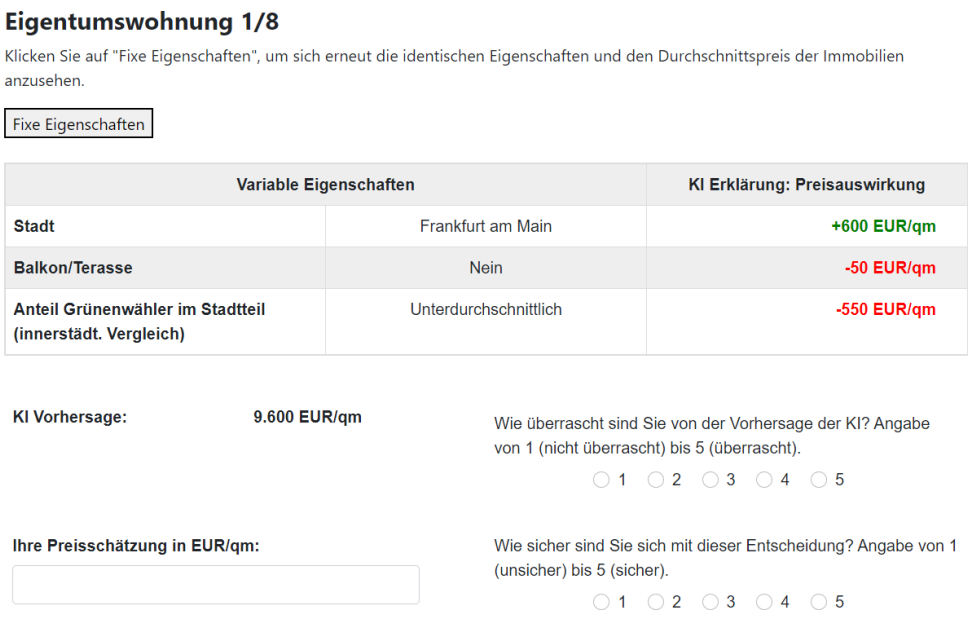
How confident are you about this decision? Indication from 1 (uncertain) to 5 (certain).

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

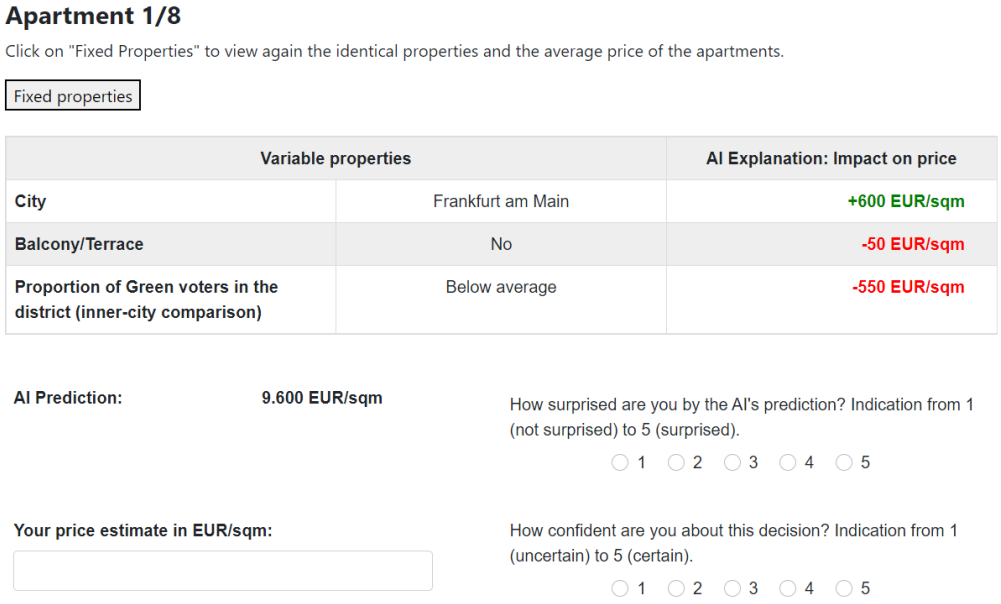
(b) English translation

Figure 9 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.



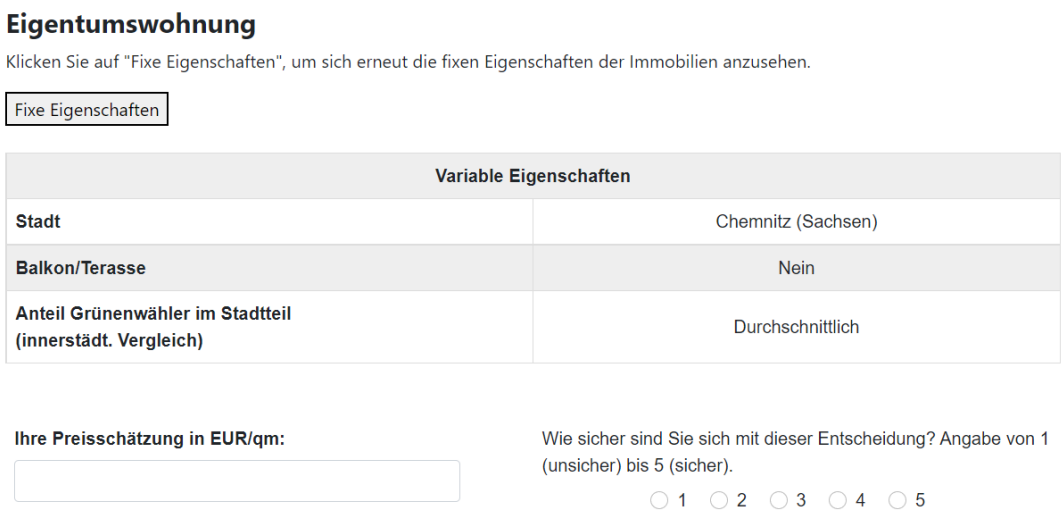
(a) Original



(b) English translation

Figure 10 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

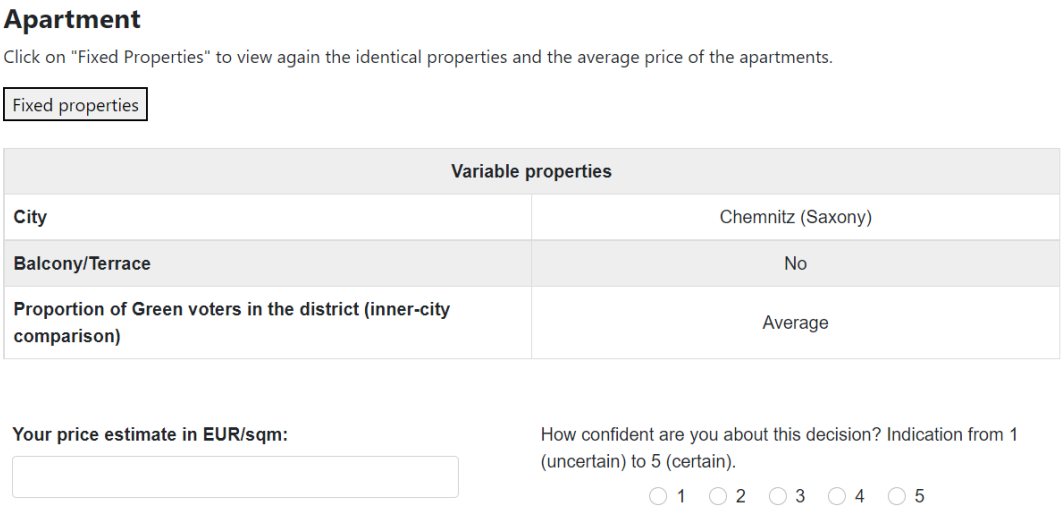
Variable Eigenschaften	
Stadt	Chemnitz (Sachsen)
Balkon/Terasse	Nein
Anteil Grünenwähler im Stadtteil (innerstädt. Vergleich)	Durchschnittlich

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 11 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 3.

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 3 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 4 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 4 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 5 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 5 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***	-0.037***	0.001
	(0.011)	(0.010)	(0.015)
Openness	0.025***	0.027***	0.002
	(0.008)	(0.008)	(0.012)
Conscien.	0.004	0.018**	0.014
	(0.008)	(0.008)	(0.012)
Agreeabln.	0.082***	0.083***	0.000
	(0.010)	(0.010)	(0.014)
Neuroticism	-0.031***	-0.009	0.022
	(0.010)	(0.010)	(0.015)
Extrav.	0.028***	0.033***	0.006
	(0.009)	(0.009)	(0.013)
Patience	0.031***	0.037***	0.006
	(0.008)	(0.008)	(0.011)
Younger sibl.	-0.005	-0.018**	-0.013
	(0.009)	(0.008)	(0.012)
Older sibl.	0.027***	0.025***	-0.001
	(0.008)	(0.008)	(0.012)
Gender (Male)	-0.038***	-0.015	0.023
	(0.009)	(0.010)	(0.014)
N	3060	3010	6070
p	0.000	0.000	0.000
R^2	0.385	0.446	0.416

Table 5 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 6 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 β_1 estimates for baseline participants' decisions in Stages I and II, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows β_2 estimates for baseline participants, measuring weight changes driven by the provision of explanations. Finally, column (6) shows DiD estimates β_4, i.e., isolated explanation-driven weight changes.

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

Table 6 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 estimates 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 6 using a logit model to demonstrate the similarity. Table 7 shows estimated marginal effects at the mean. A comparison of Tables 6 and 7 reveals that the estimates are almost identical.

	Baseline		Treatment		(5)	(6)
Dep. variable:	(1)	(2)	(3)	(4)	Prediction	Explanation
Investing	Stage I	Stage II	Stage I	Stage II	Effect	Effect
Competit.	-0.072***	-0.045***	-0.072***	-0.156***	0.031	-0.116***
	(0.019)	(0.014)	(0.021)	(0.016)	(0.021)	(0.034)
Openness	0.036**	0.022**	0.048***	0.026**	-0.017	0.002
	(0.014)	(0.011)	(0.016)	(0.011)	(0.017)	(0.028)
Agreeabln.	0.127***	0.097***	0.149***	0.011	-0.035*	-0.094***
	(0.017)	(0.014)	(0.020)	(0.012)	(0.020)	(0.033)
Extrav.	0.047***	0.020*	0.067***	0.035**	-0.031*	0.058
	(0.014)	(0.012)	(0.016)	(0.016)	(0.019)	(0.041)
Patience	0.054***	0.015	0.075***	0.082***	-0.037**	0.044*
	(0.013)	(0.012)	(0.015)	(0.013)	(0.016)	(0.026)
Older sibl.	0.041***	-0.001	0.045***	0.020**	-0.046***	0.010
	(0.013)	(0.009)	(0.015)	(0.009)	(0.017)	(0.026)
Repayment pred.		0.311***		0.243***		-0.074**
		(0.019)		(0.017)		(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 7 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 8 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 LIME × 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: Investing in Stage II	(1)
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 8 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 estimated represent LIME × 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 9 reports results from fixed-effects regression according to model (1), setting $s = 3$. Different columns show results for different subsamples of our data.

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

Table 9 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 β_1 estimates for baseline participants' decisions in Stages I and III, respectively. Columns (3) and (4) do so for treatment participants. Column (5) shows β_2 estimates for baseline participants, measuring weight changes driven by the provision of explanations. Finally, column (6) shows DiD estimates β_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 10 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 (α_1)	0.006	-0.013	-0.032	0.033	-0.015	-0.016
	(0.020)	(0.016)	(0.020)	(0.027)	(0.020)	(0.026)
Very high	-0.082***	-0.037**	-0.112***	0.011	-0.009	0.021
Competit. (α_2)	(0.027)	(0.017)	(0.028)	(0.035)	(0.021)	(0.037)
XAI × Very high	-0.004	-0.109***	-0.023	-0.024	-0.149***	-0.089***
Competit. (α_3)	(0.026)	(0.019)	(0.026)	(0.031)	(0.024)	(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 10 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 15. 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.

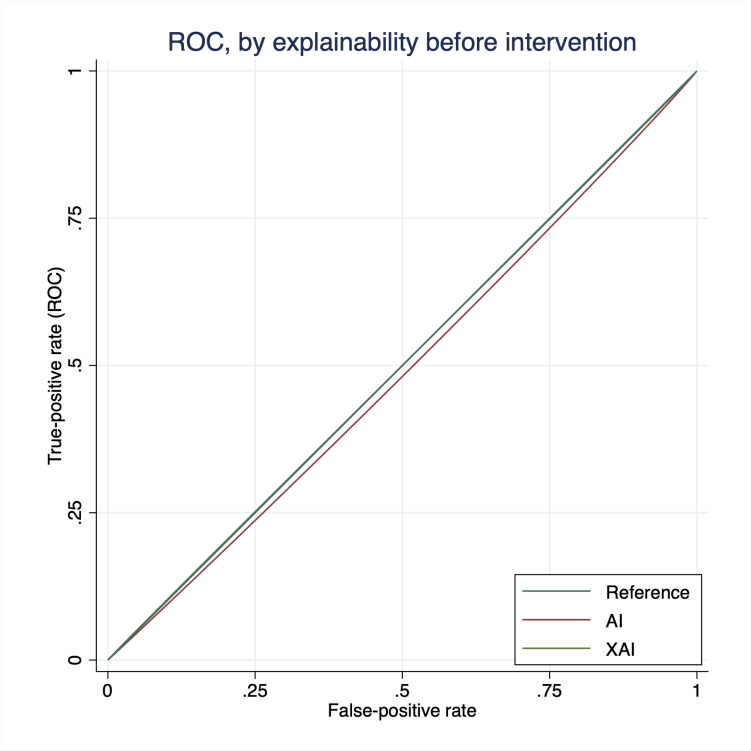


Figure 12 ROC prior to the treatment intervention.

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.

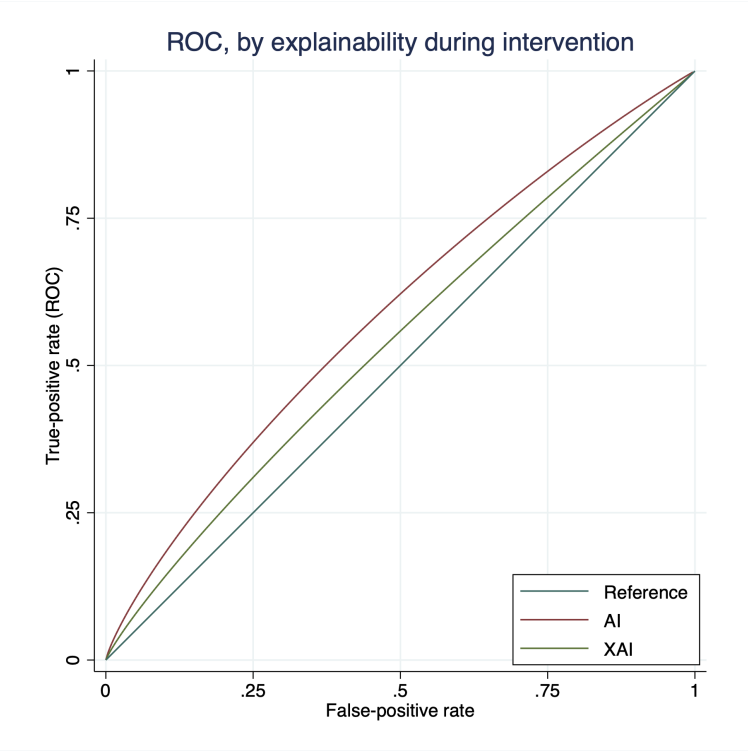


Figure 13 ROC during the treatment intervention.

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.

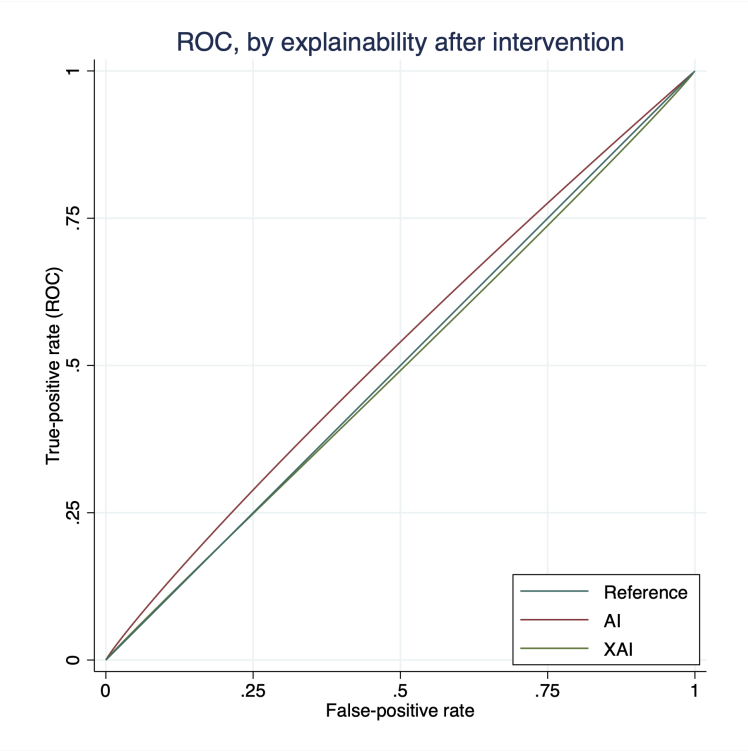


Figure 14 ROC after the treatment intervention.

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 interaction 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.

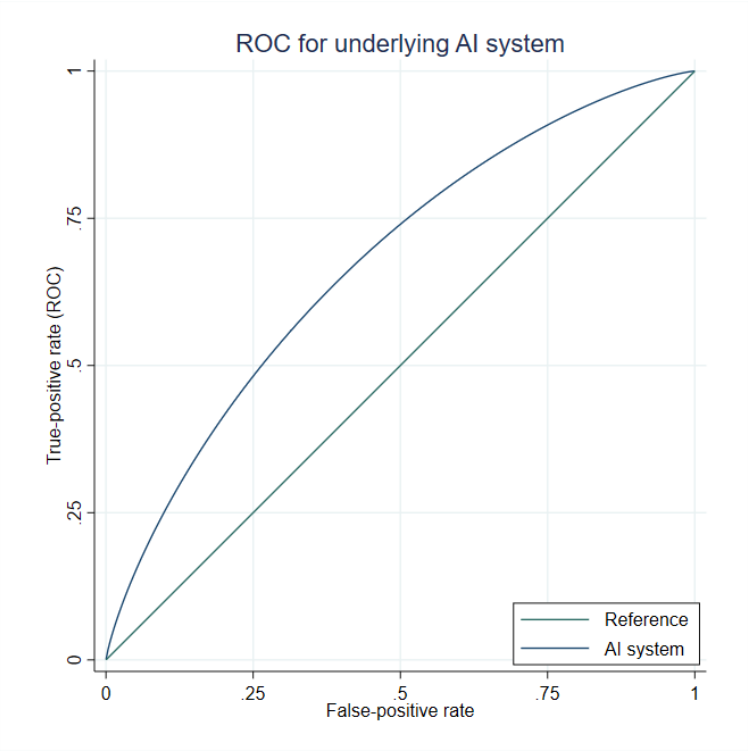


Figure 15 ROC after the treatment intervention.

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 15 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 16 shows the share of investors who selected it among the three traits to see for their investment decision.[4] 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.

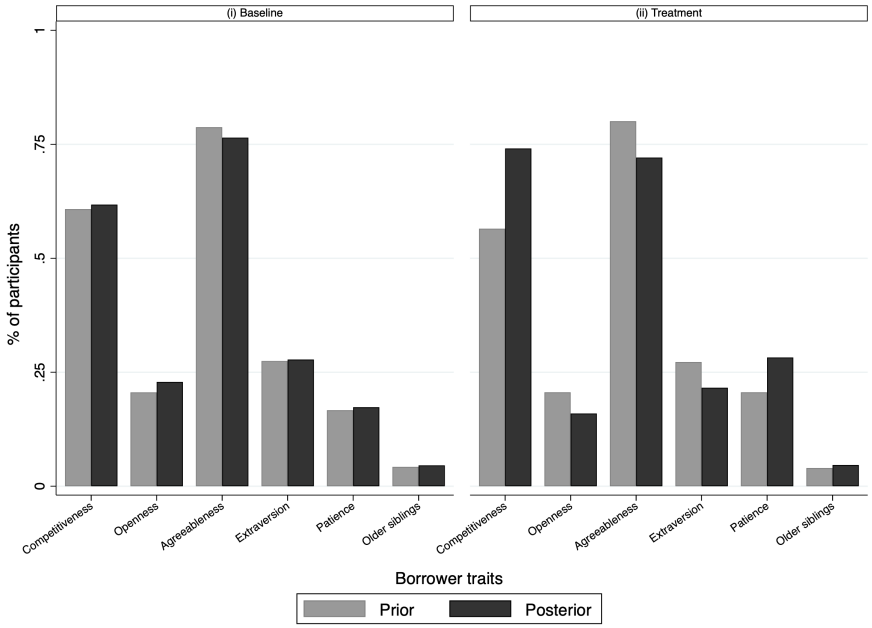


Figure 16 Preferences over observing borrower traits

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 16 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 11 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

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

Table 11 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 11). 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 11). 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 11). Considering prior and posterior preferences to see a borrower's *Agreeableness*, we do not find a significant explanation effect (see column (3) in Table 11). 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 12, 13, and 14 replicate the analysis reported in Table 6, 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.

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

Table 12 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 * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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

Table 13 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 * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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

Table 14 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 15, 16, and 17 replicate the analysis reported in Table 9, 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.

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

Table 15 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 * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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

Table 16 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 * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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

Table 17 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 18, 19, and 20 replicate the analysis reported in Table 10, 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)	(2)	(3)	(4)	(5)	(6)
	Stage I	Stage II	Stage III	Stage I	Stage II	Stage III
XAI (α_1)	0.012	-0.012	-0.027	0.041	-0.01	-0.009
	(0.020)	(0.016)	(0.020)	(0.027)	(0.020)	(0.026)
Very high	-0.077***	-0.043**	-0.107***	0.008	-0.012	0.02
Competit. (α_2)	(0.027)	(0.017)	(0.028)	(0.035)	(0.021)	(0.037)
XAI × Very high	-0.005	-0.108***	-0.025	-0.019	-0.147***	-0.086**
Competit. (α_3)	(0.026)	(0.019)	(0.027)	(0.032)	(0.024)	(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 18 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)	(2)	(3)	(4)	(5)	(6)
	Stage I	Stage II	Stage III	Stage I	Stage II	Stage III
XAI (α_1)	0.006	-0.011	-0.032*	0.035	-0.012	-0.015
	(0.019)	(0.024)	(0.019)	(0.025)	(0.018)	(0.024)
Very high	-0.102***	-0.034**	-0.131***	0.007	-0.01	0.005
Competit. (α_2)	(0.027)	(0.017)	(0.028)	(0.035)	(0.022)	(0.037)
XAI × Very high	-0.002	-0.113***	-0.023	-0.021	-0.155***	-0.088***
Competit. (α_3)	(0.026)	(0.019)	(0.026)	(0.031)	(0.024)	(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 19 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)	(2)	(3)	(4)	(5)	(6)
	Stage I	Stage II	Stage III	Stage I	Stage II	Stage III
XAI (α_1)	0.012	-0.008	-0.028	0.043*	-0.006	-0.009
	(0.019)	(0.015)	(0.019)	(0.025)	(0.018)	(0.025)
Very high	-0.097***	-0.04**	-0.127***	-0.01	-0.004	0.003
Competit. (α_2)	(0.028)	(0.018)	(0.028)	(0.035)	(0.023)	(0.038)
XAI × Very high	-0.003	-0.112***	-0.025	-0.016	-0.153***	-0.085***
Competit. (α_3)	(0.026)	(0.019)	(0.027)	(0.032)	(0.024)	(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 20 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 21 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 (α_1)	51.617	-3.713
	(33.397)	(51.199)
Expl. (α_2)	20.513	135.410***
	(32.298)	(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 21 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 22 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: Posterior belief	(1) Overall	(2) Low confidence beliefs	(3) High confidence beliefs
Prior belief	0.496***	0.463***	0.748***
	(0.061)	(0.070)	(0.105)
Avg. SHAP	0.397***	0.424***	0.249***
	(0.028)	(0.037)	(0.047)
Confirm	-21.026	-51.430	131.448*
	(30.661)	(45.167)	(77.220)
Prior belief × Confirm	0.117	0.250**	-0.342*
	(0.093)	(0.112)	(0.190)
Avg. SHAP × Confirm	-0.123	-0.324**	0.231**
	(0.087)	(0.139)	(0.110)
N	708	481	222
p	0.000	0.000	0.000
R^2	0.743	0.728	0.840

Table 22 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 23 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 × AI	-298.992	-233.176	-2073.208
	(801.443)	(884.999)	(1736.720)
Balcony × Expl.	-79.185	-687.224	283.148
	(746.115)	(1567.682)	(1339.158)
Low green × AI	894.361	483.990	2575.297
	(993.437)	(1086.211)	(2197.200)
High green × AI	-390.458	717.449	-2551.901
	(1064.479)	(1019.652)	(2056.778)
Low green × Expl.	-1902.323**	495.868	-3459.922**
	(877.642)	(2091.666)	(1632.984)
High green × Expl.	1742.126*	84.433	3906.959***
	(911.520)	(1736.347)	(1312.532)
N	153	72	81
p	0.000	0.000	0.000
R^2	0.289	0.471	0.527

Table 23 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* × *XAI* and *High green* × *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.

References

Ai C, Norton EC (2003) Interaction terms in logit and probit models. *Economics letters* 80(1):123–129.

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)*.

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.

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

Gramegna A, Giudici P (2021) SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence* 4:752558.

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.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Conference on Neural Information Processing Systems (NIPS)* .

Nori H, Jenkins S, Koch P, Caruana R (2019) InterpretML: A unified framework for machine learning interpretability. *arXiv:1909.09223*.

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*.