Manipulating and Measuring Model Interpretability

FOROUGH POURSABZI-SANGDEH, Microsoft Research
DANIEL G. GOLDSTEIN, Microsoft Research
JAKE M. HOFMAN, Microsoft Research
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research

With machine learning models being increasingly used to aid decision making even in high-stakes domains, there has been a growing interest in developing interpretable models. Although many supposedly interpretable models have been proposed, there have been relatively few experimental studies investigating whether these models achieve their intended effects, such as making people more closely follow a model's predictions when it is beneficial for them to do so or enabling them to detect when a model has made a mistake. We present a sequence of pre-registered experiments ($N = 3,800$) in which we showed participants functionally identical models that varied only in two factors commonly thought to make machine learning models more or less interpretable: the number of features and the transparency of the model (i.e., whether the model internals are clear or black box). Predictably, participants who saw a clear model with few features could better simulate the model's predictions. However, we did not find that participants more closely followed its predictions. Furthermore, showing participants a clear model meant that they were *less* able to detect and correct for the model's sizable mistakes, seemingly due to information overload. These counterintuitive findings emphasize the importance of testing over intuition when developing interpretable models.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **User studies**.

Additional Key Words and Phrases: interpretability, machine-assisted decision making, human-centered machine learning

ACM Reference Format:
Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 67 pages. https://doi.org/10.1145/3411764.3445315

1 INTRODUCTION

Machine learning models are increasingly used to aid decision making in high-stakes domains, such as medical diagnosis [57], credit risk assessment [43], judicial sentencing and bail [5, 18, 48, 53], and hiring [66]. Machine learning models also influence people's decisions about what news articles to read [3, 13, 69, 88, 104], what movies to watch [10], what music to listen to [77], what clothes to buy [20], and even who to date [90]. In all of these settings, decision making is a collaboration between people and models, where models make predictions and people can choose whether to follow these predictions or to override them.

CHI '21, May 8-13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445315

There are many reasons why following a machine learning model's predictions may be advantageous, chief among them being improved accuracy. Indeed, there have been many studies showing that models are often more accurate than people. A meta-analysis from twenty years ago, reviewing work from some seventy years ago, found that models were more accurate than people in a variety of domains [24, 41], and the gap has widened since then as models have become more accurate [53]. Although following a model's predictions should enable people to make faster and more consistent decisions, there are also scenarios in which following a model's predictions can be disadvantageous, most notably when those predictions are incorrect.

That said, people are resistant to using models to aid their decision making [8, 23, 25, 86]. There are many reasons for this: First, people may feel that they do not understand models, including what information they rely on and how this information is being used. For instance, in a study of machine learning use in the public sector [106], several practitioners noted challenges in getting organizational buy-in for the use of machine learning-based systems without the ability to explain those systems' internals. Second, people may feel that models do not rely on the right information [24, p. 151] or that they do not use information in the right ways [106]. Third, people may be worried that models might behave in ways that are unfair [22, 33, 54, 80, 91, 106, 109]. Concerns about fairness are often exacerbated by the first two reasons.

In response to these concerns, a prolific line of research has emerged that focuses on the *interpretability* of machine learning models. There are two main approaches to developing supposedly interpretable models. First, because there is evidence that simple models with clear internals can be as accurate as more complex, black-box models in some domains [6, 23, 34, 48, 92, 95], one approach is to create simple, clear models such as point systems that can be memorized [48, 102] or generalized additive models that facilitate visualizing the impact of each feature on the model's predictions [15, 72, 73]. The hope is that these models will be easier for people to understand and use.

The second approach is to provide post-hoc explanations for potentially complex, black-box models. Threads of research that focus on this approach look at how to explain individual predictions by learning simple local approximations of a model around particular data points [63, 74, 89], training simple models to mimic more complex ones [64, 98], estimating the influence of training data points [56], describing the change to an input data point that would change a model's prediction for it [93, 103, 108, 113], and visualizing model predictions or properties [58, 112].

But despite this progress, there is still no consensus about how to define, quantify, or measure the interpretability of a machine learning model [30], raising the following question: What is interpretability and how can we determine whether one model is more interpretable than another? Different notions of interpretability, such as simplicity, transparency, simulatability, and trustworthiness, are often conflated [68]. This problem is exacerbated by the fact that machine learning models have many different types of stakeholders and these stakeholders may have different needs in different scenarios [44, 99, 100]. The approach that works best for a regulator who wants to understand why a particular person was denied a loan may be different to the approach that works best for a data scientist debugging a machine learning model or for a CEO using a model to make a high-stakes decision. Moreover, regardless of how interpretability is defined, quantified, or measured, there is very little scientific evidence demonstrating that a) people are better able to understand interpretable models, b) people more closely follow the predictions of interpretable models when it is beneficial for them to do so, and c) people are better able to detect when an interpretable model has made a mistake, enabling them to override its prediction.

We take the perspective that the lack of consensus around defining, quantifying, or measuring interpretability, as well as the lack of scientific evidence for its benefits, stem from the fact that interpretability is not something that can be directly manipulated or measured. Rather, the

interpretability of a model is a latent—and fundamentally human—property that can be influenced by different *manipulable factors* (such as the number of features, the complexity of the model, the transparency of the model, or even the user interface) and that impacts different *measurable (human) outcomes* (such as people's abilities to simulate the model's predictions, the extent to which people follow the model's predictions when it is beneficial for them to do so, or people's abilities to detect when the model has made a mistake). Different factors may influence different outcomes in different ways. As such, we argue that to understand interpretability, it is necessary to directly manipulate different factors and measure their effects. What is or is not interpretable must be defined by people's behavior, not by what appeals to intuition [65, 78, 79, 110].

Drawing on this perspective, we present a sequence of pre-registered experiments ($N = 3,800$) in which we varied factors commonly thought to make machine learning models more interpretable and measured their effects on people's behavior. Based on a structured review of the literature on interpretability, we focused on two factors—the number of features and the transparency of the model (i.e., whether the model internals are clear or black box)—and investigated how these factors affected three measurable outcomes:

(1) **How well can people simulate a model's predictions?**
(2) **To what extent do people follow a model's predictions when it is beneficial for them to do so?**
(3) **How well can people detect when a model has made a mistake and correct for it?**

We found that people can better simulate the predictions of a clear model with few features compared to the predictions of a clear model with more features or the predictions of a black-box model. However, contrary to our expectations, we did not find a significant improvement in the extent to which people follow the predictions of a clear model with few features when it is beneficial for them to do so compared to the predictions of a black-box model with more features. We also found that using a clear model hampers people's abilities to detect when the model had made a mistake. All three of these findings are based on highly-powered, pre-registered experiments with multiple, representative stimuli. Our latter two findings are notable and surprising because they contradict common intuition about interpretability.

1.1 Domain

We focused our experiments on the domain of real-estate valuation, in which machine learning models are used to predict the selling prices of properties.[1] In each experiment, participants were asked to predict the prices of apartments in a single neighborhood in New York City with the help of a machine learning model. We conducted our experiments on laypeople, as laypeople represent one type of stakeholder that might potentially use or be affected by machine learning models. We chose the domain of real-estate valuation because many people have considered purchasing a property, making the setting both familiar and potentially interesting to participants. Each apartment was represented in terms of eight features: number of bedrooms, number of bathrooms, square footage, total rooms, days on the market, maintenance fee, distance from the subway, and distance from a school. All participants saw the same set of apartments (i.e., the same feature values) and, crucially, the same predicted selling price for each apartment, which came from either a two-feature or an eight-feature linear regression model. To achieve this, the models were constrained to make the same predictions for the apartments we used, as we describe below in Section 3.1. What varied between the experimental conditions was therefore *only the presentation of the model*: whether it was presented as using two- or eight-features and whether the model internals were clear or black

[1]The Zestimate prices on the website Zillow may be a familiar example of these predictions.

box. As a result, any observed differences in participants' behavior could be attributed entirely to the presentation of the model—a key feature of our experimental design.

Because of our decision to vary only the presentation of the model, each participant had access to all eight feature values for each apartment, regardless of the experimental condition to which they were assigned. This meant that some participants (those who were shown an eight-feature model) had access to the same information as the model, while others (those who were shown a two-feature model) had access to more information than the model. This scenario has a rich history in the decision-making literature, where it is called the "broken leg problem" [24, p. 151], based on an anecdote in which a model that is very good at predicting weekly attendance at the movies should be ignored if it is known that someone has a broken femur with a full length cast. This scenario is also often encountered in practice: Table 1 in appendix A contains a number of instances from the literature in which people have access to more information than a model, meaning that they can use their knowledge of this additional information as a reason to override the model's predictions.

1.2 Overview of experiments

In our first experiment, we showed participants a sequence of twelve apartments. The first ten apartments had typical configurations (i.e., typical combinations of feature values), whereas the last two had unusual configurations (such as three bathrooms squeezed into 726 square feet). For each apartment, participants were first shown its configuration (i.e., feature values) alongside the model (whose internals were either clear or black box) and were asked to guess what the model would predict for the apartment's selling price. They were then shown the model's prediction and asked for their own prediction of the apartment's selling price.

We hypothesized that participants who were shown the clear, two-feature model would better simulate the model's predictions [60] and would more closely follow its predictions when it was beneficial for them to do so. We also hypothesized that participants assigned to different experimental conditions would be differently able to detect and correct for the model's sizable mistakes on the apartments with unusual configurations. We note that here and throughout the rest of paper, when we refer to detecting and correcting a model's mistakes, we are specifically referring to whether participants notice that the model has made an inaccurate prediction and provide a more accurate prediction themselves; doing so does not necessarily imply that they understand why the model made the mistake.

As expected, we found that participants who saw the clear model with two features could better simulate the model's predictions. However, we did not find that participants more closely followed its predictions when it was beneficial for them to do so. Moreover, participants' predictions were generally less accurate than the models'—a familiar finding in the literature on the predictions of people versus computational systems [24, 40, 41]. Furthermore, and contrary to our intuition when designing the experiment, participants who were shown a clear model were *less* able to detect and correct for the model's sizable mistakes on the apartments with unusual configurations compared to participants who were shown a black-box model. To account for these unexpected findings, we designed and ran three additional experiments.

In our second experiment, we scaled down the apartments' selling prices and maintenance fees to match median prices in the U.S. in order to determine whether the findings from our first experiment were merely an artifact of New York City's high prices. Reassuringly, with scaled-down selling prices and maintenance fees, the findings from our first experiment replicated quite closely.

Our third experiment used *weight of advice*—a measure commonly used in the literature on advice-taking [36, 114] and subsequently used in the context of computational decision making by Logg [70, 71]—as an alternative way to measure the extent to which people follow a model's predictions.

Here too, we found no significant differences in the extent to which participants followed the predictions of the model when it was beneficial for them to do so between the experimental conditions. Surprisingly, and contrary to our findings from the previous two experiments, we did *not* find that participants who were shown a clear model were less able to detect and correct for the model's sizable mistakes.

We conjectured two possible reasons for this finding. First, in all three experiments, participants may have anchored on the prediction visible to them when making their own final prediction of an apartment's selling price. However, the possible anchor values in the first two experiments were different to that in the third: in the third experiment, participants saw their own initial prediction of each apartment's selling price when making their final prediction, whereas in the first two experiments, participants saw their simulation of the model's prediction. In the first two experiments, participants who were shown the clear, two-feature model could better simulate the model's predictions compared to participants assigned to the other experimental conditions, and might therefore have anchored on higher selling prices for the apartments with unusual configurations, in line with the model's predictions. Additionally, participants who were shown a clear model may have been overwhelmed by the amount of detail in front of them, causing them to be less likely to notice the unusual apartment configurations when making their own predictions. This effect may have been less prominent in our third experiment because participants made their initial predictions before being shown the model.

This motivated our fourth and final experiment. For this experiment, we returned to the design of our first experiment, but removed the simulation step and varied whether or not participants were shown an "outlier focus" message highlighting the apartments with unusual configurations as possible outliers. We found that participants who were shown a clear model and no outlier focus message were less able to detect and correct for the model's sizable mistakes, as in our first two experiments. In contrast, this difference disappeared for participants who were shown an outlier focus message, in line with the findings from our third experiment. The findings from our fourth experiment are therefore consistent with the possible explanation outlined above.

In light of this, we then conducted some additional post-hoc analyses of the data from our first two experiments, finding that participants who were shown a clear, eight feature model (i.e., the model presentation with the most information) were worse at simulating the model's predictions [60] and followed its predictions less closely compared to participants assigned to the other experimental conditions. We also found that these participants' predictions of the apartments' selling prices were less accurate. These findings, which we present along with our findings from the fourth experiment, are also consistent with the explanation above.

To summarize, via a sequence of pre-registered experiments involving several thousand participants, we found that two factors commonly thought to make machine learning models more interpretable often have negligible effects on people's behavior and, in some cases, even have detrimental effects. Contrary to the intuition that models with clear internals can only improve people's decisions, our findings suggest otherwise. Taken together, our findings emphasize the importance of testing over intuition when developing interpretable models.

In the next section, we further situate our experiments in the literature from the machine learning, human–computer interaction, and decision-making communities. In the subsequent four sections, we describe our experiments and present our findings in detail. We then conclude by discussing limitations of and possible extensions to our work, as well as implications for designing user interfaces that facilitate effective collaborations between people and models.

2 RELATED WORK

Although there has been a recent surge of research in the machine learning community on techniques for achieving interpretability [15, 48, 56, 61, 63, 64, 74, 89, 93, 98, 102, 103, 108], there have been relatively fewer studies of how factors commonly thought to make machine learning models more interpretable affect people's behavior. Perhaps most closely related and contemporaneous to our work, Lage et al. [60] used controlled experiments involving laypeople to investigate how the complexity of a model affects its simulatability, focusing on decision sets. They found that the number of cognitive chunks and the model size both affect people's abilities to simulate a model's predictions. Other researchers have conducted user studies in order to understand people's use of specific tools or methods. For example, Huysmans et al. [45] studied the effects of presenting people with models that are traditionally thought to be more interpretable (such as decision tables and binary decision trees) on people's accuracies and their stated confidences in completing a task; Lim et al. [67] studied the effects of different types of explanations (such as probing a machine learning model about why it made a particular prediction or why it did not make a different prediction) on laypeople's understandings of and trust in a model; Rader et al. [87] studied the effects of different ways of explaining Facebook's News Feed algorithm on people's understandings of how the algorithm works and their ability to evaluate the correctness of the algorithm's output; Cheng et al. [17] studied the effects of different design and interface choices for presenting explanations on people's understandings of and trust in computational decisions; Binns et al. [11] and Dodge et al. [29] studied the effects of different types of explanations on people's perceptions of a model's fairness; and Kaur et al. [49] studied data scientists' use of two specific interpretability tools (the InterpretML [81] implementation of generalized additive models and the SHAP Python package), finding that data scientists over-trust and misuse these tools. More commonly, machine learning researchers include small-scale user studies to evaluate their own proposed techniques. For example, Lakkaraju et al. [62] ran a user study comparing 47 students' understandings of decision boundaries corresponding to interpretable decision sets versus Bayesian decision lists, while Ribeiro et al. [89] ran experiments to investigate whether laypeople are able to use local interpretable model-agnostic explanations to choose which of two classifiers is better, to perform feature engineering, and to identify classifier irregularities.

Within the human–computer interaction community, there is a longstanding practice of taking a user-centered perspective and acknowledging that people are active participants who form their own mental models of how computational systems work [32, 47, 82]. Bellotti and Edwards [9] argued for design principles that support intelligibility, so that systems "represent to their users what they know, how they know it, and what they are doing about it." Similarly, Glass et al. [37] demonstrated empirically that transparency around how complex adaptive agents work improves people's trust in those agents. Stumpf et al. [97] were among the first researchers to address the role of mental models when people interact with machine learning-based systems. They conducted a series of experiments to study the benefits of allowing rich interactions between people and systems, assessing whether different types of explanations would better enable people to form useful mental models. They and others [55] also found that people become more willing to use computational systems when they are given the opportunity to review and potentially modify the systems, even when the modifications have no effects [104]. Kulesza et al. [59] studied several ways in which intelligent agents might explain themselves to stakeholders. They showed that completeness of explanations is more important than soundness in accurately shaping mental models, but that people lose trust in a system when soundness is too low.

Another line of human–computer interaction research that relates to forming mental models focuses on sensemaking [85, 94]. Sensemaking refers to the process by which people collect and

organize information and acquire "situation awareness" (i.e., build a mental model of the knowledge and data at hand). In the context of our work, sensemaking relates to people's understandings of machine learning models, while situation awareness facilitates insight and enables people to make intelligent decisions. Sensemaking research often involves designing tools to support rich interactions among people [39] or between people and computational systems. Sensemaking processes are likely operating in our experiments when participants examine the model to simulate its predictions. Sensemaking processes may also be at play when participants detect and correct for the model's sizable mistakes, though we do not collect cognitive process measures to investigate their reasoning directly. We do, however, test a hypothesis about information overload that rests on an assumption about interference in the information-intake process.

Finally, there is considerable research related to our experiments in the decision-making literature. To date, much of this work has focused on people's aversion [8, 23, 25] or proclivity [71] to trust computational decision-making aids, and ways to increase this trust [26]. Other relevant decision-making work has endorsed the creation of simple or "improper" linear models that bear a strong resemblance to the models that we used in our experiments [23, 34, 38, 48]. Although decision-making researchers have tested the accuracies of these models in simulations, there have been far fewer tests of these models when used by people to aid their decision making. In our experiments, which we describe starting in the next section, we extend this line of research by taking a slightly different approach and asking how presentation differences in functionally identical models—specifically, differences in two factors thought to make machine learning models more interpretable—affect people's behavior.

3 EXPERIMENT 1: PREDICTING APARTMENT SELLING PRICES

Our first experiment was designed to measure the effects of the number of features and the transparency of the model (i.e., whether the model internals are clear or black box) on three measurable outcomes that our literature review revealed to be often associated with interpretability: laypeople's abilities to simulate a model's predictions, the extent to which laypeople follow a model's predictions when it is beneficial for them to do so, and laypeople's abilities to detect when a model has made a mistake. Before running the experiment, we posited and pre-registered three hypotheses, stated informally below:[2]

H1. **Simulation.** Participants will better simulate the predictions of a clear model with few features.

H2. **Deviation.** For typical data points, participants will more closely follow (i.e., deviate less from) the predictions of a clear model with few features when it is beneficial for them to do so compared to the predictions of a black-box model with more features.

H3. **Detection of mistakes.** Participants assigned to different experimental conditions will be differently able to detect and correct for the model's sizable mistakes on unusual data points.

We tested the first hypothesis by showing each participant an apartment's configuration (i.e., feature values), asking them to guess what the model would predict for the apartment's selling price, and then comparing this prediction to the model's prediction. A small difference between these two quantities, which we refer to as the participant's simulation error, indicates that the participant could better simulate the model's prediction.

For the second hypothesis, we measured the extent to which each participant deviated from the model's predictions by, for each of the apartments with typical configurations, showing them the model's prediction, asking them for their own prediction of the apartment's selling price, and measuring the difference between these two quantities.

[2]Pre-registered hypotheses for this experiment are available at https://aspredicted.org/xy5s6.pdf.

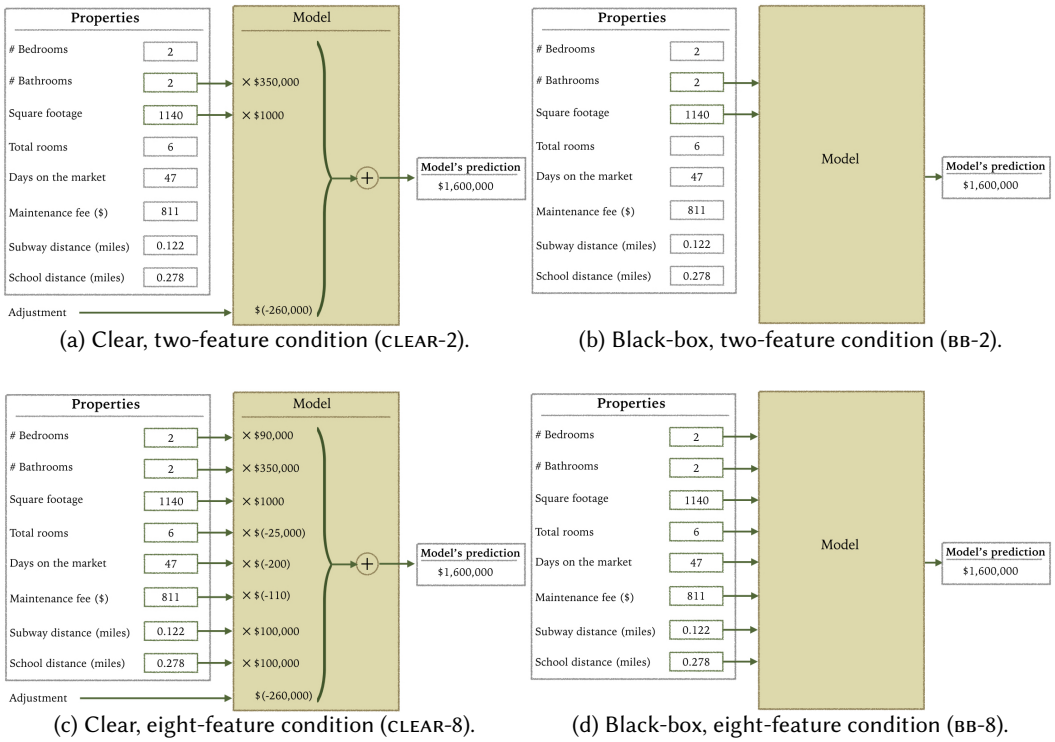
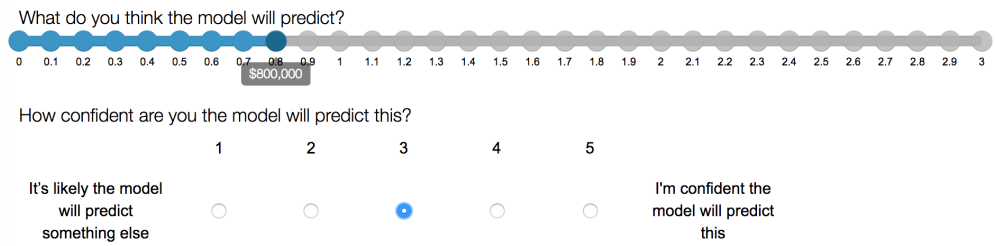


Fig. 1. The four primary experimental conditions. In the conditions in the top row, the model used two features; in the conditions in the bottom row, the model used eight. In the conditions on the left, the model internals were clear; in the conditions on the right, the model internals were black box.

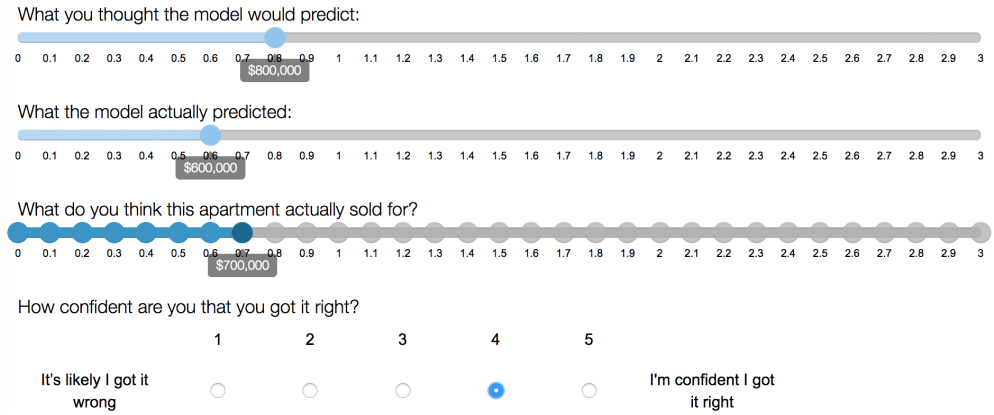
We used the same measure for the third hypothesis, but focused on only the apartments with unusual configurations. Specifically, we said that a participant was able to detect and correct for the model's sizable mistakes if we saw large deviations between the model's predictions and their own predictions for the apartments with unusual configurations. We note that this does not necessarily imply that they understood why the model made the mistakes. We did not pre-register any directional hypotheses about which experimental conditions would result in participants being more or less able to detect and correct for the model's sizable mistakes. On the one hand, if a participant better understands the model, they may be better equipped to detect and correct for its overly high predictions. On the other hand, a participant may more closely follow the model's predictions if they better understand it. For this reason, we pre-registered our third hypothesis to be bi-directional, but we note that our intuition at the time was that participants who were shown a clear model would be better able to detect and correct for its sizable mistakes, compared to participants who were shown a black-box model.

We additionally pre-registered our intent to analyze participants' prediction errors (i.e., how far their own predictions of the apartments' selling prices were from the actual selling prices), but again refrained from pre-registering any directional hypotheses.



(a) Step 1: Participants were asked to guess what the model would predict and state their confidence in this guess.

(b) Step 2: Participants were asked to state their confidence in the model's prediction.



(c) Step 3: Participants were asked for their own prediction and to state their confidence in this prediction.

Fig. 2. Part of the testing phase from our first experiment.

3.1 Experimental design

As we explained in Section 1, we asked participants to predict the selling prices of apartments in a single neighborhood in New York City with the help of a machine learning model. To do this, we used a 2×2 design:

- Participants were randomly assigned to see either a two-feature model (number of bathrooms and square footage—the two most predictive features) or an eight-feature model.
- Participants were randomly assigned to either see a clear model (i.e., a linear regression model with visible coefficients) or a black-box model.

We additionally included a baseline condition in which there was no model available to participants.

All participants saw the same set of apartments (i.e., the same feature values). The models were constrained to make the same predictions for these apartments, so participants saw the same model predictions regardless of the experimental condition to which they were assigned (see Appendix C). Furthermore, the accuracies of the models were nearly identical, as described below. What varied between the experimental conditions was therefore *only the presentation of the model*. This was a key feature of our experimental design that enabled us to run tightly controlled experiments.

Screenshots from each of the four primary experimental conditions (i.e., each experimental condition in our 2×2 design, but not the baseline condition) are shown in Figure 1. We note that each participant had access to all eight feature values for each apartment, regardless of the experimental condition to which they were assigned.

We ran the experiment on Amazon Mechanical Turk using psiTurk [42], an open-source platform for designing online experiments. Multiple studies have shown that data from high-reputation Turkers is comparable to data from commercial panels and university pools when assessing outcomes such as attentiveness, honesty, and effort [14, 16, 21, 75, 84].[3]

We recruited 1,250 participants, all located in the U.S., with approval ratings greater than 97%. We randomly assigned participants to the experimental conditions (CLEAR-2, $N = 248$; CLEAR-8, $N = 247$; BB-2, $N = 247$; BB-8, $N = 256$; and NO-MODEL, $N = 252$). Each participant received a flat payment of $2.50. The experiment was approved by our institutional review board.

Participants were first shown detailed instructions, including, in the conditions involving clear models, a simple English description of the corresponding two- or eight-feature model (see Appendix B). To ensure that participants understood these instructions, each participant was required to answer a multiple choice question about the number of features used by the model before proceeding with the experiment, which consisted of two phases.

The *training phase* familiarized participants with both the domain (i.e., real-estate valuation) and the model's predictions. Participants were shown ten apartments in a random order. In the four primary experimental conditions, participants were shown the model's prediction of each apartment's selling price, asked for their own prediction of the apartment's selling price, and then shown the apartment's actual selling price. In the baseline condition (i.e., no model), participants were asked to predict the selling price for each apartment and then shown its actual selling price.

In the *testing phase*, participants were shown twelve apartments that they had not previously seen. The order of the first ten apartments was randomized, while the remaining two apartments always appeared last, for the reasons described below. In the four primary experimental conditions,

[3] Although recent research has shown that Turkers may manipulate their demographic information so as to be included in studies [96], we only screened Turkers using two criteria, both of which are enforced by the platform and would require some effort to manipulate: country and approval rating. Moreover, even if some participants had manipulated their information, it would have had only a minor effect on our findings because we were not trying to estimate quantities relating to the entire population of participants, but were instead trying to detect differences between the experimental conditions. That is, the proportion of Turkers with manipulated information would be, on average, the same for each condition, thereby permitting valid measurement of randomly assigned treatment effects.

participants were asked to guess what the model would predict for each apartment's selling price (i.e., simulate the model's prediction) and to state their confidence in this guess on a five-point scale (see Figure 2a). They were then shown the model's prediction and asked to state their confidence in that prediction (see Figure 2b). Finally, they were asked for their own prediction of the apartment's selling price and to state their confidence in this prediction (Figure 2c). In the baseline condition, participants just were asked to predict the selling price of each apartment and to state their confidence in this prediction.

We selected the apartments from a data set of apartments sold between 2013 and 2015 on the Upper West Side of New York City, taken from StreetEasy.com, a popular real-estate website. To create the models, we first fit a two-feature linear regression model (i.e., estimated the model's coefficients) using this data set and rounded the coefficients for readability.[4] To ensure that the models were as similar as possible, we fixed the intercept and the coefficients for number of bathrooms and square footage in an eight-feature model to match those of the two-feature model, and then fit the model (i.e., estimated the model's coefficients for the remaining six features) and followed the same rounding procedure as with the two-feature model. The rounded coefficients for both models are shown in Figure 1. The models explain 82% of the variance in the apartments' selling prices. When presenting the models' predictions to participants, we rounded each prediction to the nearest $100,000.

All participants saw the same set of apartments (i.e., the same feature values) because randomizing the selection would have introduced additional noise and reduced the power of the experiment, making it harder to spot differences between the experimental conditions. To enable comparisons between the experimental conditions, the ten apartments in the training phase and the first ten apartments in the testing phase were selected from the apartments in our data set for which the rounded predictions of the two- and eight-feature models were the same. We selected the apartments to cover a representative range of model prediction errors (i.e., how far the models' predictions were from the apartments' actual selling prices). We provide details of the apartments' configurations (i.e., feature values) and our selection procedure in Appendix C.

We used the last two apartments in the testing phase to test our third hypothesis—namely, that participants assigned to different experimental conditions will be differently able to detect and correct for the model's sizable mistakes on unusual data points. Ideally, we would have used two apartments with unusual configurations for which both models made the same sizable mistakes. Unfortunately, there were no such apartments in our data set, so we selected (in one case) and synthetically generated (in the other) two apartments to test different aspects of our hypothesis. These apartments' configurations exploited the models' large coefficient ($350,000) for number of bathrooms. The first apartment ("apartment 11") was a one-bedroom, two-bathroom apartment (selected from our data set) for which both models made overly high, but different, predictions. As a result, comparisons between the conditions involving the two-feature model and the conditions involving the eight-feature model were therefore impossible, although we were able to analyze participants' prediction errors because these did not rely on the models' predictions. The second apartment ("apartment 12") was a synthetically generated one-bedroom, three-bathroom, 726-square-foot apartment for which both models made the same overly high prediction, allowing us to make comparisons between all four primary experimental conditions, but ruling out analyses of participants' prediction errors because the apartment did not have an actual selling price. We emphasize that even though it did not have an actual selling price, we are confident that it would have been overpriced by the models because of its three bathrooms squeezed into only 726 square

[4]For each coefficient, we found a round number that was within one quarter of a standard error of the estimated coefficient.

feet. Apartments 11 and 12 were always shown last to avoid the phenomenon in which people trust a model less after seeing it make a mistake [25].

3.2 Findings

Having run the experiment, we compared participants' behavior across the conditions.[5] Doing so required us to compare multiple responses (i.e., data about multiple apartments) from multiple participants, which was complicated by possible correlations among each participant's responses. For example, some participants might have consistently overestimated the apartments' selling prices regardless of the condition to which they were assigned, while others might have consistently provided underestimates. We addressed this by fitting a mixed-effects model for each measurable outcome of interest to capture differences between conditions while controlling for participant-level effects—a standard approach for analyzing repeated measure experimental designs [7]. We derived all statistical tests from these models. Bar plots and mean outcomes in the density plots correspond to average values (± one standard error) by condition from the fitted mixed-effects models. To test our hypotheses, we ran contrasts and calculated degrees of freedom, test statistics, and p-values under these models. Unless otherwise noted, all plots and statistical tests correspond to just the first ten apartments from the testing phase.[6]

Our findings are as follows:

H1. **Simulation.** We defined each participant's simulation error for each apartment to be $|m - u_m|$—i.e., the absolute difference between the model's prediction of the apartment's selling price m and the participant's guess for the model's prediction u_m. Figure 3a contains density plots for participants' mean simulation errors. Participants assigned to the condition involving the clear, two-feature model had, on average, lower simulation errors compared to participants assigned to the other primary experimental conditions ($t\,(994) = -12.06$, $p < 0.001$). This means that, as hypothesized, participants could better simulate the predictions of the clear, two-feature model.

H2. **Deviation.** We defined each participant's deviation from the model's prediction of each apartment's selling price to be $|m - u_a|$—i.e., the absolute difference between the model's prediction of the apartment's selling price m and the participant's own prediction of the apartment's selling price u_a. Figure 3b shows that contrary to our second hypothesis, we did not find a significant difference in the extent to which participants followed the predictions of the clear, two-feature model when it was beneficial for them to do so compared to the predictions of the black-box, eight-feature model ($t\,(994) = 0.67$, $p = 0.5$).

H3. **Detection of mistakes.** As explained above, we used the last two apartments in the testing phase (apartments 11 and 12) to test our third hypothesis. The models made overly high predictions for these apartments because of their unusual configurations. For both apartments, participants assigned to the four primary experimental conditions predicted higher selling prices compared to participants assigned to the baseline condition (i.e., no model). We suspect that this is because participants anchored on the models' predictions. For apartment 11, we found no significant differences in participants' deviations from the model's predictions between the four primary

[5]For each of our experiments, we report all sample sizes, conditions, data exclusions, and measures for the main analyses that were described in our pre-registration documents. We determined the sample size for our first experiment based on estimates from a small pilot experiment, which enabled us to detect a difference of at least $50,000 in deviation between the condition involving the clear, two-feature model and the condition involving the black-box, eight-feature model with 80% power. For our subsequent experiments, however, we adjusted the sample size to target a power of 80% or more. We provide full distributions of participants' responses in Appendix E and details of our statistical tests in Appendix F.

[6]All the data and code needed to reproduce our results are available at https://github.com/Foroughp/Manipulating-and-Measuring-Model-Interpretability.

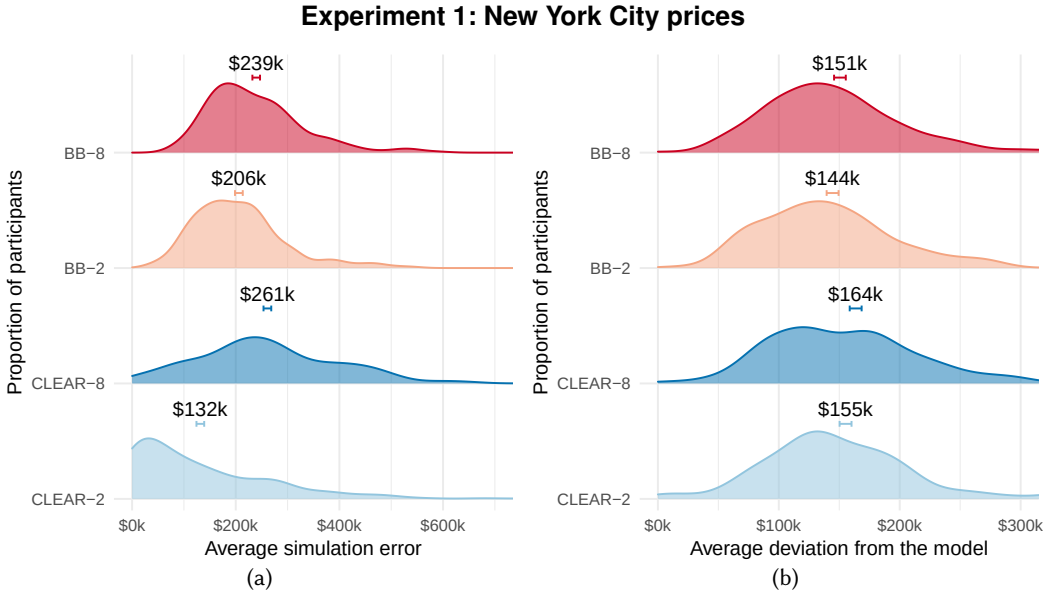


Fig. 3. Results from our first experiment: density plots for participants' (a) mean simulation errors and (b) mean deviations from the model's predictions. Numbers in each subplot indicate average values over all participants in the corresponding condition, while error bars indicate one standard error.

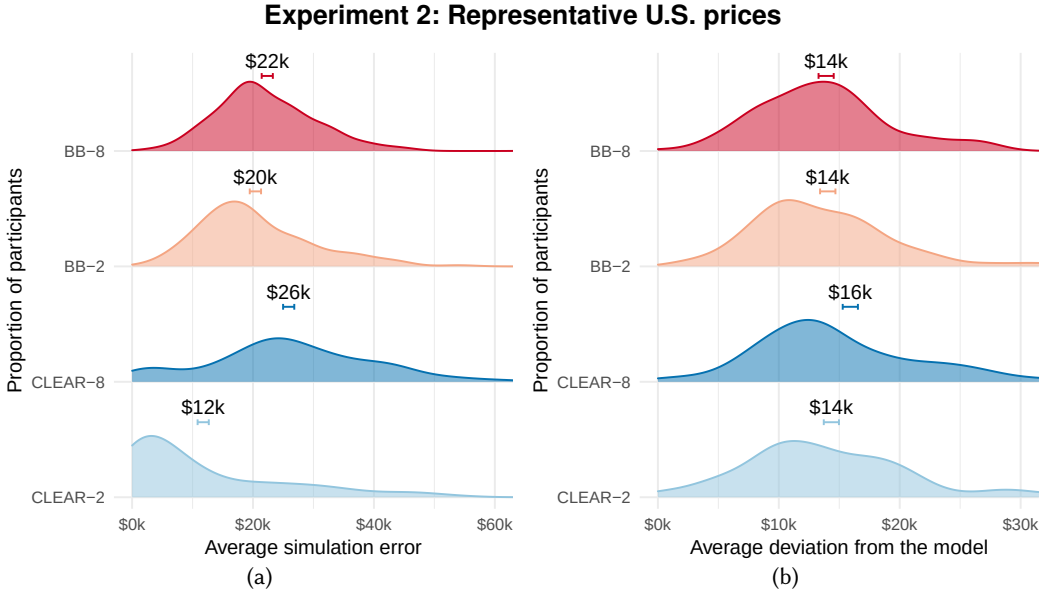


Fig. 4. Results from our second experiment, which replicate the findings from our first experiment.

experimental conditions ($F\,(3, 994) = 1.03$, $p = 0.379$ under a one-way ANOVA). In other words, we found that participants assigned to different experimental conditions were similarly able to detect and correct for the model's overly high prediction for apartment 11. For apartment 12, a

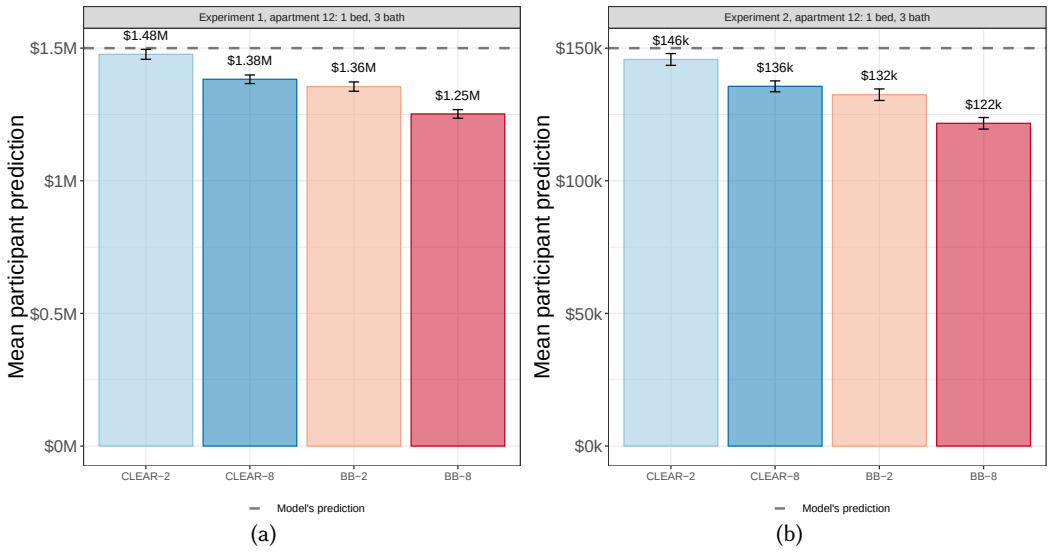


Fig. 5. Participants' mean predictions of apartment 12's selling price in (a) our first experiment and (b) our second experiment. Horizontal lines indicate the models' predictions and error bars indicate one standard error.

one-way ANOVA revealed a significant difference in participants' deviations from the model's predictions between the four primary experimental conditions ($F\,(3, 994) = 4.42$, $p = 0.004$). Participants assigned to the conditions involving clear models deviated from the models' prediction, on average, less compared to participants assigned to the conditions involving black-box models ($F\,(1, 994) = 8.81$, $p = 0.003$ for the main effect of the transparency of the model, see Figure 5a). This finding contradicts our intuition when designing the experiment, which was that participants who were shown a clear model would be better able to detect and correct for its sizable mistakes compared to participants who were shown a black-box model. We explore this finding in more detail in Section 6.

We also conducted some post-hoc analyses. First, we analyzed participants' stated confidences in the models' predictions for each apartment. Although we did not pre-register a hypothesis about this, we found an interesting difference between participants' stated confidences and their revealed behavior. Specifically, even though participants assigned to the condition involving the clear, two-feature model stated that they were more confident in the model's predictions compared to participants assigned to the condition involving the black-box, eight-feature model (on average, a difference of .25 on a five-point scale from "I'm confident the model got it wrong" to "I'm confident the model got it right," ($t\,(994) = 4.27$, $p < 0.001$)), their behavior did not reflect this. We found no significant differences in the extent to which participants followed the model's predictions between the four primary experimental conditions.

Our second post-hoc analysis involved participants' prediction errors. We defined each participant's prediction error for each apartment to be $|a - u_a|$—i.e., the absolute difference between the apartment's actual selling price a, and the participant's own prediction of the apartment's selling price u_a. A one-way ANOVA did not reveal any significant differences in participants' prediction errors between the four primary experimental conditions ($F\,(3, 994) = 2.43$, $p = 0.06$).

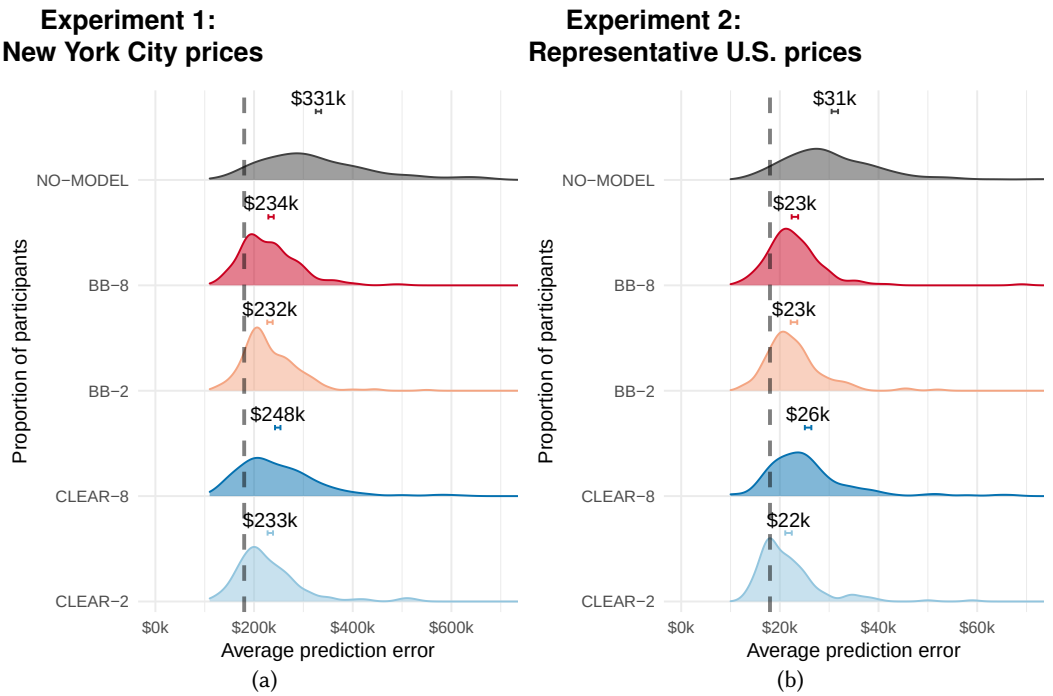


Fig. 6. Density plots for participants' mean prediction errors in our first experiment (left) and in our second experiment (right). Numbers in each subplot indicate average values over all participants in the corresponding condition, while error bars indicate one standard error. Vertical lines indicate the model's mean prediction error.

As shown in Figure 6a, we also found that for the apartments with typical configurations, participants assigned to the four primary experimental conditions had, on average, higher prediction errors than the model, but lower prediction errors than participants assigned to the baseline condition ($t\,(1245) = 15.28$, $p < 0.001$ for the comparison of the baseline condition with the four primary experimental conditions), suggesting that using a model was advantageous. In contrast, participants' prediction errors for apartment 11 revealed the opposite pattern: participants assigned to the four primary experimental conditions had, on average, lower prediction errors than the model but higher prediction errors than participants assigned to the baseline condition ($t\,(1245) = -7.99$, $p < 0.001$). Using the model helped participants more accurately predict the selling prices of the apartments with typical configurations, but hindered them when predicting the selling price of an apartment with an unusual configuration for which the model had made a sizable mistake.

Additionally, we found that using a clear model further hampered participants when making their own predictions about apartment 11's selling price: participants who were shown a clear model made, on average, less accurate predictions compared to participants who were shown a black-box model ($F\,(1, 994) = 31.98$, $p < 0.001$ for the main effect of the transparency of the model under a two-way ANOVA).

To summarize, as hypothesized, we found that participants who were shown the clear, two-feature model could better simulate the model's predictions. However, we did not find that they followed its predictions more closely when it would have been beneficial for them to do so. We also found that, contrary to our intuition, participants who were shown a clear model were less able to

detect and correct for the model's sizable mistakes on unusual data points. Finally, we found no differences in participants' prediction errors between the four primary experimental conditions. We also found that that using a model is advantageous, but that participants would have been better off (i.e., had lower prediction errors) had they followed the model's predictions for the apartments with typical configurations.

4 EXPERIMENT 2: REPRESENTATIVE U.S. PRICES

One potential concern about our first experiment is that participants' lack of familiarity with New York City's unusually high prices might have influenced the extent to which they followed the models or their abilities to detect and correct for the models' sizable mistakes. Our second experiment was therefore designed as a robustness check targeted at this concern.

In this experiment, we replicated our first experiment but with the apartments' selling prices and maintenance fees scaled down to match median prices in the U.S. Before running this experiment we again posited and pre-registered three hypotheses.[7] The first two hypotheses (H4 and H5) were identical to the first two hypotheses from our first experiment. However, we made the third hypothesis (H6) more precise than the third hypothesis from our first experiment to reflect the findings from that experiment, as well as our findings from a small pilot experiment with scaled-down prices. This hypothesis is stated informally below:

H6. **Detection of mistakes.** Participants who are shown a clear model will be less able to detect and correct for the model's sizable mistakes on unusual data points, and this effect will be more prominent for more unusual data points (i.e., for apartment 12 compared with apartment 11).

4.1 Experimental design

We first scaled down the apartments' selling prices and maintenance fees by a factor of ten. To account for this change, we also scaled down all coefficients (except for the coefficient for maintenance fee) by a factor of ten. Apart from the description of the neighborhood from which the apartments were selected, the experimental design was unchanged from our first experiment. We again ran the experiment on Amazon Mechanical Turk using psiTurk. We excluded Turkers who had participated in our first experiment, and recruited 750 new participants, all of whom satisfied the screening criteria from our first experiment. We randomly assigned participants to the experimental conditions (CLEAR-2, $N = 150$; CLEAR-8, $N = 150$; BB-2, $N = 147$; BB-8, $N = 151$; and NO-MODEL, $N = 152$). Again, each participant received a flat payment of $2.50.

4.2 Findings

The findings from our first experiment replicated quite closely.

H4. **Simulation.** As hypothesized, and as shown in Figure 4a, participants assigned to the condition involving the clear, two-feature model had, on average, lower simulation errors compared to participants assigned to the other primary experimental conditions ($t\,(594) = -10.41$, $p < 0.001$). This is in line with the findings from our first experiment.

H5. **Deviation.** Also in line with the findings from our first experiment, but contrary to our hypothesis, we found no significant difference in the extent to which participants followed the predictions of the clear two-feature model when it was beneficial for them to do so compared to the predictions of the black-box, eight-feature model ($t\,(594) = 0.49$, $p = 0.626$, see Figure 4b).

H6. **Detection of mistakes.** For apartment 11, although a one-way ANOVA revealed a significant difference between the four primary experimental conditions ($F\,(3, 594) = 3.00$, $p = 0.03$), as was

[7] Pre-registered hypotheses for this experiment are available at https://aspredicted.org/3bv8i.pdf.

the case in our first experiment, we found no significant differences between the conditions involving clear models and the conditions involving black-box models ($t\,(594) = -1.82$, $p = 0.069$), perhaps because apartment 11's configuration was not sufficiently unusual. For apartment 12, in line with the findings from our first experiment, and as hypothesized, a one-way ANOVA revealed a significant difference between the four primary experimental conditions ($F\,(3, 594) = 7.96, p < 0.001$). In particular, participants assigned to conditions involving clear models followed the model's prediction, on average, more closely than participants assigned to conditions involving black-box models, indicating that they were *less* able to detect and correct for the model's overly high prediction, thereby resulting in an even worse final prediction for the apartment's selling price ($t\,(594) = -4.16$, $p < 0.001$, see Figure 5b).

We again conducted some post-hoc analyses. In contrast to the findings from our first experiment, we found no significant difference in participants' stated confidences between the condition involving the clear, two-feature model and the condition involving the black-box, eight-feature model ($t\,(594) = 1.03$, $p = 0.303$). We note that the effect size of the difference in our first experiment was small (Cohen's d of 0.23) and even smaller in our second experiment (Cohen's d of 0.07), which was identical to the first, except for prices. We also note that there was no significant difference in the extent to which participants followed the model's predictions when it was beneficial for them to do so in either experiment.

We also analyzed participants' prediction errors. Here, a one-way ANOVA did reveal a significant difference in participants' prediction errors between the four primary experimental conditions ($F\,(3, 594) = 8.60$, $p < 0.001$). That said, the maximum pairwise difference in prediction error between the four primary experimental conditions is not large ($4,000 or roughly 3% of the average selling price, which was $120,000).

In line with the findings from our first experiment, we found that for the apartments with typical configurations, participants assigned to the four primary experimental conditions had, on average, higher prediction errors than the model, but lower prediction errors than participants assigned to the baseline condition ($t\,(745) = 10.62$, $p < 0.001$). Again, we found the opposite pattern for apartment 11: participants assigned to the four primary experimental conditions had, on average, lower prediction errors than the model, but higher prediction errors than participants assigned to the baseline condition ($t\,(745) = -6.41$, $p < 0.001$). We also found, in line with the findings from our first experiment, that using a clear model further hampered participants when making their own predictions about apartment 11's selling price: participants who were shown a clear model made less accurate predictions compared to participants who were shown a black-box model ($F\,(1, 594) = 7.16$, $p = 0.008$ for the main effect of the transparency of the model under a two-way ANOVA).

To summarize, the main findings from our second experiment closely replicate the findings from our first experiment, suggesting that New York City's unusually high prices did not influence participants' behavior. In both experiments, we found that participants who were shown a clear, two-feature model could better simulate the model's predictions. However, they did not follow the model's predictions more closely when it was beneficial for them to do so. They were also less able to detect and correct for the model's sizable mistakes on unusual data points.

5 EXPERIMENT 3: WEIGHT OF ADVICE

In our first two experiments, we did not find a significant difference in the extent to which participants followed the predictions of the clear, two-feature model when it was beneficial for them to do so compared to the predictions of the black-box, eight-feature model, measured in terms of the absolute difference between the model's prediction and the participant's own prediction for each of the apartments with typical configurations. Because this finding was contrary to our hypotheses,

we wondered whether using an alternative way to measure the extent to which people follow a model's predictions would yield a different finding. In our third experiment, we therefore used *weight of advice*—a measure commonly used in the literature on advice-taking [36, 70, 114].

Weight of advice quantifies the extent to which people update their beliefs (e.g., their own predictions made *before* seeing a model's predictions) toward any advice they are given (e.g., the model's predictions). In the context of our first two experiments, each participant's weight of advice for each apartment is defined as $\frac{|u_a^{(2)} - u_a^{(1)}|}{|m - u_a^{(1)}|}$, where m is the model's prediction of the apartment's selling price , $u_a^{(1)}$ is the participant's initial prediction of the apartment's selling price before seeing m, and $u_a^{(2)}$ is the participant's final prediction of the apartment's selling price after seeing m. Weight of advice is equal to 1 if the participant's final prediction matches the model's prediction and equal to 0.5 if the participant averages their initial prediction and the model's prediction.

To understand the benefits of weight of advice, consider the scenario in which a participant's final prediction $u_a^{(2)}$ is close to the model's prediction m. There are two reasons why this might happen. On the one hand, it could be the case that the participant's initial prediction $u_a^{(1)}$ was far from m and they made a significant update to their initial prediction after seeing the model's prediction. On the other hand, it could be the case that the participant's initial prediction $u_a^{(1)}$ was already close to m, so they did not update their prediction at all after seeing the model's prediction. The absolute difference between the model's prediction m and the participant's final prediction $u_a^{(2)}$ does not distinguish between these two cases. In contrast, weight of advice does—i.e., it will be high in the first case and low in the second.

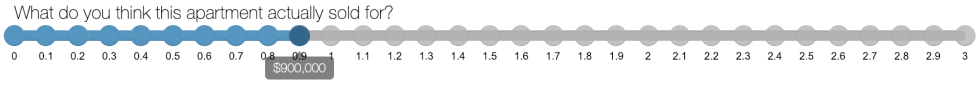
We additionally used our third experiment to check whether participants' behavior would be different if they were told that the predictions were made by a "human expert" instead of a model. Previous studies have examined this question from different perspectives with differing results [25, 27, 28, 31, 83]. Most closely related to our experiment the work of Logg [70, 71], which showed that when people have no information about the quality of the predictions they are shown, they follow the predictions that appear to come from a computational system more closely than those that appear to come from a person. We were curious to see whether this would also be the the case when people were given a chance to assess the quality of the predictions before deciding how closely to follow them.

The details of our hypotheses for this experiment are provided in Appendix D.

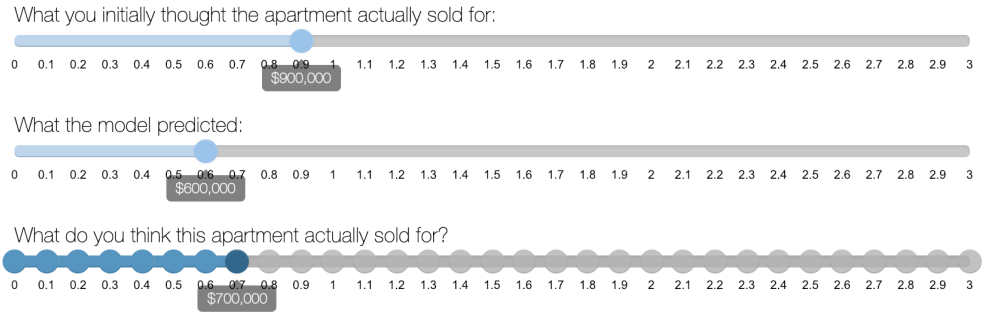
5.1 Experimental design

For this experiment, we returned to the original New York City prices and used the same four primary experimental conditions as in the first two experiments. However, we also added a new condition, in which participants saw exactly the same information as in the condition involving the black-box, eight feature model, but with the model labeled as "Human Expert" instead of "Model." We did not include a baseline condition because the most natural baseline would have been to simply ask each participant for their own prediction of each apartment's selling price, which was already the first half of this experiment's testing phase, as described below.

As before, we ran the experiment on Amazon Mechanical Turk using psiTurk. We excluded Turkers who had participated in our first two experiments, and recruited 1,000 new participants, all of whom satisfied the screening criteria from our first two experiments. However, when analyzing the data, we excluded the responses from one participant who reported technical difficulties with the experiment. We randomly assigned participants to the experimental conditions (CLEAR-2, $N = 202$; CLEAR-8, $N = 200$; BB-2, $N = 202$; BB-8, $N = 198$; and EXPERT, $N = 197$). For this experiment, each participant received a flat payment of $1.50.



(a) First half: Participants were asked for their initial prediction of the apartment's selling price.



(b) Second half: Participants were shown the model's prediction and asked to update their prediction.

Fig. 7. Part of the testing phase from our third experiment.

We asked participants to predict the selling prices of the same apartments that we used in our first two experiments. However, we slightly modified the testing phase so that we could calculate weight of advice. In particular, each participant was asked for two predictions of each apartment's selling price: an initial prediction before being shown the model's prediction and a final prediction after being shown the model's prediction. To keep the length of the experiment reasonable, we did not ask participants to guess what the model would predict for each apartment's selling price.

We also designed the experiment so as to elicit each participant's initial predictions for all twelve apartments before showing them the model. This is because we ran a small experiment in which participants were first shown an apartment's configuration (i.e., feature values) and asked for their prediction of its selling price. They were then shown the model's prediction—and the model itself, whose internals were either clear or black box—and asked to update their prediction before moving on to the next apartment. We found that participants assigned to the condition involving the clear, two-feature model made initial predictions that were closer to the model's predictions—even though they had not seen the model's prediction when making their initial prediction—compared to participants assigned to the other primary experimental conditions ($t\ (239) = -3.42$, $p < 0.001$). We suspect that this is because the clear, two-feature model was easiest for participants to simulate. As a result, participants may have more easily internalized the model's coefficients when making their final prediction for an apartment and then used them to make their initial predictions for subsequent apartments. Although this kind of behavior is often beneficial, here it posed a threat to the validity of our experiment: for us to be able to compare participants' weight of advice between different experimental conditions, a participant's initial predictions should not be influenced by the condition to which they were assigned.

As in our first two experiments, participants were first shown detailed instructions (which, this time, intentionally did not include any information about the model or "human expert"), before proceeding with the experiment, which consisted of two phases. In the (short) training phase, participants were shown three apartments in a random order. For each one, they were asked for their prediction of the apartment's selling price and shown the actual selling price. The testing phase consisted of two halves. In the first half, participants were shown another twelve apartments. The order of all twelve apartments was randomized. Participants were asked for their initial prediction of each apartment's selling price (see Figure 7a). In the second half, participants were first introduced to the model or "human expert" before revisiting the twelve apartments (see Figure 7b). The order of the first ten apartments was randomized, while the remaining two (apartments 11 and 12) always appeared last, as in the first two experiments. For each apartment, participants were first reminded of their initial prediction, then shown the model or expert's prediction, and only then asked to make their final prediction of the apartment's selling price. To simplify the experiment, we did not ask participants to state their confidence in either their or the model's predictions.

5.2 Findings

We briefly summarize our findings here and provide full details in Appendix D. This experiment confirmed our findings from the first two experiments about the extent to which participants followed the predictions of the clear, two-feature model when it was beneficial for them to do so compared to the predictions of the black-box, eight-feature model. We again found no significant difference in how closely participants followed the predictions of the clear, two-feature model compared to the predictions of the black-box, eight-feature model—this time measured in terms of weight of advice, as well as in terms of the absolute difference between the model's prediction and the participant's final prediction for each of the apartments with typical configurations.

We also found that participants followed the predictions of the "human expert" no more closely than they followed the predictions of the black-box models. We suspect that the difference between this finding and those of Logg [70, 71] is due to participants' increasing experience with the model or "human expert" over the course of our experiment.

Finally, in contrast to the findings from our first two experiments, we did *not* find that participants assigned to the conditions involving clear models were less able to detect and correct for the model's overly high predictions for either apartment 11 or apartment 12. This last finding motivated our final experiment, which we describe in the next section.

6 EXPERIMENT 4: OUTLIER FOCUS AND DETECTION OF MISTAKES

Contrary to our intuition when designing the first two experiments, participants who were shown a clear model in those experiments were less able to detect and correct for the model's sizable mistakes on apartments with unusual configurations compared to participants assigned to conditions involving black-box models (see Figures 5a and 5b). In our third experiment, in seeming contradiction, we found no such difference between the conditions involving clear models and the conditions involving black-box models. In this section, we propose a possible explanation for these findings and then support it with a final experiment. The explanation rests on two reasons, which we outline below.

First, in all three experiments, participants who were shown a clear model may have been overwhelmed by the amount of detail in front of them—i.e., they may have experienced information overload[8] [2, 46, 50]—causing them to be less likely to notice the unusual apartment configurations

[8]We emphasize that we are referring to visual information overload that affects attention to items on a display [19], not cognitive load in working memory, which has also been shown to be related to interpretability [1, 60].

when making their own predictions. We conjecture that this effect may have been less pronounced in our third experiment, though, because participants were asked for their initial predictions for all twelve apartments' selling prices before being introduced to the model. In turn, this may have meant that they paid greater attention to each each apartment's configuration—unusual or not.

Second, in all three experiments, participants may have anchored on the prediction visible to them when making their own final prediction of an apartment's selling price [26, 101]. However, the possible anchor values differed between the experiments: In the first two experiments, participants made their final prediction of each apartment's selling price while seeing their *simulation of the model's prediction* (see Figure 2c). In contrast, in the third experiment, participants made their final prediction of each apartment's selling price while seeing their *own initial prediction of the apartment's selling price* (see Figure 7b).

Furthermore, in the first two experiments, the anchor values differed between the experimental conditions because they were influenced by the model involved. Participants assigned to the condition involving the clear, two-feature model could better simulate the model compared to participants assigned to the other experimental conditions (see Figures 3a and 4a). However, if the model has overpriced an apartment, then better simulating it might cause participants to anchor on a selling price that is too high. On top of that, because clear models reveal more information, participants may have been even less likely to notice the unusual apartment configurations due to information overload. In contrast, participants assigned to the conditions involving black-box models were not able to simulate the model so well and, perhaps undistracted by what was in front of them, may have been more likely to notice the unusual apartment configurations. Interestingly, participants assigned to the conditions involving black-box models apparently (incorrectly) assumed that the model would take the unusual apartment configurations into account and therefore made lower guesses for the model's predictions. In other words, in the first two experiments, participants assigned to the conditions involving black-box models could have had two things working in their favor: they were less likely to be overwhelmed by the amount of detail in front of them and they may have anchored on their lower guesses for the model's predictions.

We designed our fourth experiment to test this possible explanation. As we describe below, this experiment removed the potential for anchoring and measured the effect of an "outlier focus" message highlighting the apartments with unusual configurations as possible outliers (see Figure 8). In our previous experiments, the number of features did not appear to have a strong effect on participants' abilities to detect and correct for the model's sizable mistakes, so we used only the two-feature linear regression model in this experiment.

Before running the experiment, we posited and pre-registered three hypotheses, stated informally below:[9]

H11. **Outlier focus.** Participants that see an outlier focus message and participants that don't see an outlier focus message will be differently able to detect and correct for the model's sizable mistakes on unusual data points.

H12. **Transparency (clear vs. black box) and no outlier focus.** When they are not shown an outlier focus message, participants who are shown a clear model and participants who are shown a black-box model will be differently able to detect and correct for the model's sizable mistakes on unusual data points.

H13. **Transparency (clear vs. black box) and outlier focus.** When they are shown an outlier focus message, participants who are shown a clear model and participants who are shown a black-box model will be differently able to detect and correct for the model's sizable mistakes on unusual data points.

[9]Pre-registered hypotheses for this experiment are available at https://aspredicted.org/5xy8y.pdf.

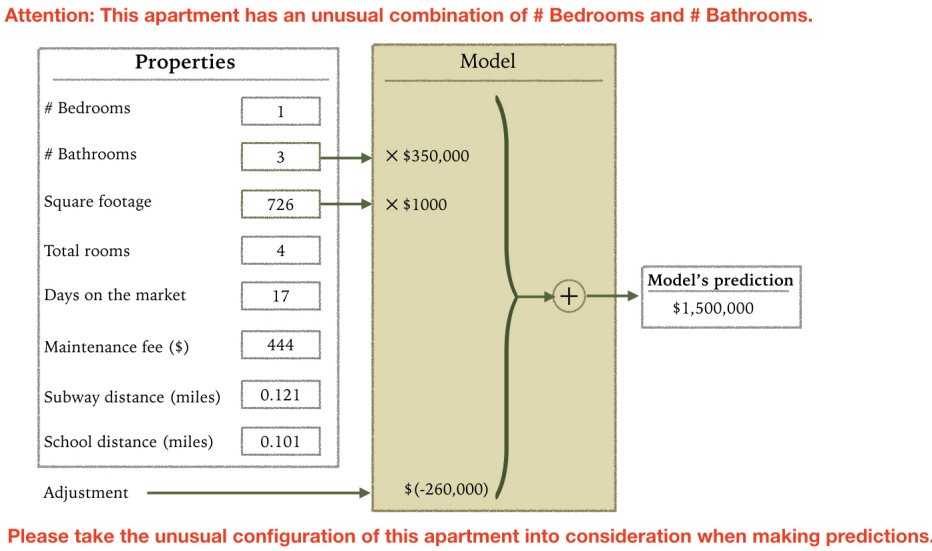


Fig. 8. Apartment 6 in the conditions involving an outlier focus message in our fourth experiment.

6.1 Experimental design

Similar to the design of our first experiment, we asked participants to predict the selling prices of apartments in New York City with the help of a machine learning model. We used a 2×2 design:

- Participants were randomly assigned to see either a clear model (i.e., a linear regression model with visible coefficients) or a black-box model.
- Participants were randomly assigned to either see an outlier focus message highlighting the apartments with unusual configurations as possible outliers or to not see such a message.

We again ran the experiment on Amazon Mechanical Turk using psiTurk. We excluded Turkers who had participated in our first three experiments, and recruited 800 new participants, all of whom satisfied the screening criteria from our first three experiments. We randomly assigned participants to the experimental conditions (CLEAR-FOCUS, $N = 202$; CLEAR-NO-FOCUS, $N = 195$; BB-FOCUS, $N = 201$; and BB-NO-FOCUS, $N = 202$). Each participant received a flat payment of $1.00.

To keep the experiment short, we limited the training phase and the first portion of the testing phase to only five of the original ten apartments previously used in each phase. We used three apartments for the second portion of the testing phase: two synthetically generated apartments with unusual configurations, with an apartment with a typical configuration (one bedroom, one bathroom, 788 square feet) in between. The first synthetically generated apartment ("apartment 6") was apartment 12 from our previous experiments (a one-bedroom, three-bathroom, 726-square-foot apartment). The second synthetically generated apartment ("apartment 8") had an even more unusual configuration (one bedroom, three bathrooms, 350 square feet) and, like apartment 6, was overpriced by the model. The order of the two synthetically generated apartments was randomized, while the apartment with the typical configuration ("apartment 7") was always shown in the middle.

The training phase was the same as in the first two experiments (except with fewer apartments). In the testing phase, participants were shown eight apartments, described above. For each apartment, participants were first shown the model's prediction and asked to state their confidence in that prediction. They were then asked for their own prediction of the apartment's selling price and to

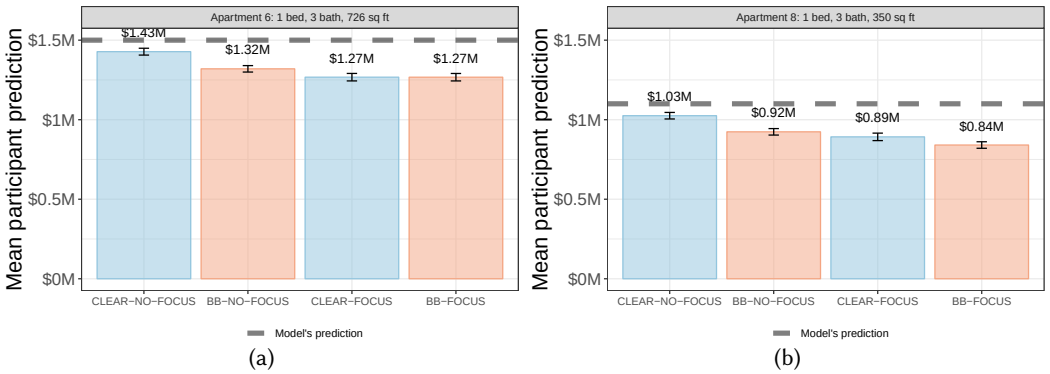


Fig. 9. Results from our fourth experiment: participants' mean predictions of the selling prices for the apartments with unusual configurations: (a) apartment 6 and (b) apartment 8. Horizontal lines indicate the model's predictions and error bars indicate one standard error.

state their confidence in this prediction. To remove the potential for anchoring, participants were not asked to guess what the model would predict for each apartment.

6.2 Findings

Figure 9 shows participants' mean predictions of the selling prices for the apartments with unusual configurations (i.e, apartment 6 and apartment 8). To test our hypotheses, we defined each participant's deviation from the model's prediction of each apartment's selling price to be $u_a - m$, where m is the model's prediction and u_a is the participant's prediction of the apartment's selling price. We used signed difference (rather than absolute difference, as in our first two experiments) because the goal of this experiment was to study participants' abilities to detect and correct for the model's mistakes. Using signed difference enabled us to more easily tell whether a participant's deviation from the model's prediction was in the right direction.

H11. **Outlier focus.** We found that participants in conditions involving an outlier focus message deviated from the model's predictions, on average, more compared to participants who did not see an outlier focus message for both apartment 6 ($t\,(791) = -4.72$, $p < 0.001$) and apartment 8 ($t\,(795) = -5.00$, $p < 0.001$). In other words, showing participants an outlier focus message better enabled them to detect and correct for the model's sizable mistakes on the apartments with unusual configurations.

H12. **Transparency (clear vs. black box) and no outlier focus.** Participants assigned to conditions involving the clear model and no outlier focus message deviated from the model's predictions, on average, less compared to participants assigned to conditions involving the black-box model and no outlier focus message for both apartment 6 ($t\,(393) = -3.65$, $p < 0.001$) and apartment 8 ($t\,(395) = -3.51$, $p < 0.001$). In other words, in line with the findings from our first two experiments, without an outlier focus message, participants who were shown the clear model were less able to detect and correct for the model's sizable mistakes on the apartments with unusual configurations, compared to participants who were shown a black-box model.

H13. **Transparency (clear vs. black box) and outlier focus.** We found no significant difference in participants' deviations from the model's predictions between the condition involving the clear model and an outlier focus message and the condition involving the black-box model and an outlier focus message ($t\,(401) = -0.004$, $p = 0.996$ for apartment 6 and $t\,(394) = -1.64$, $p = 0.101$ for apartment 8). In other words, with an outlier focus message, participants who were shown the

clear model were similarly able to detect and correct for the model's sizable mistakes, compared to participants who were shown a black-box model. This finding suggests that an outlier focus message helps participants pay attention to information that they might otherwise miss due to information overload.

Taken together, these findings support our explanation for the difference between the findings from our first two experiments and the findings from our third. They also highlight an unintended disadvantage of using clear models—and offer a simple way to mitigate it.

In light of this, we returned to the data from our first two experiments and conducted some additional post-hoc analyses. First, we analyzed participants' prediction errors. In our first experiment, although a one-way ANOVA did not reveal a significant difference in participants' prediction errors between the four primary experimental conditions (see Section 3), a visual inspection of our results (see Figure 6a) indicates that participants assigned to the condition involving the clear, eight-feature model had, on average, higher prediction errors than participants assigned to the other primary experimental conditions. Indeed, this difference is statistically significant ($t\,(994) = 2.68$, $p = 0.007$). Of course, we note that with large sample sizes, statistical significance might not mean practical significance [76, 107]. Indeed, this seems to be the case with participants' prediction errors. For example, in Figure 6a, the maximum pairwise difference in prediction error between the four primary experimental conditions is quite small—only about \$16,000 or roughly 1% of the average selling price, which was \$1.2 million. In contrast, in Figure 3a, the maximum pairwise difference in simulation error between the four primary experimental conditions is more substantial at \$129,000.

Analyzing the data from our second experiment revealed a similar pattern. Here, a one-way ANOVA did reveal a small but significant difference in participants' prediction errors between the four primary experimental conditions (see Section 4). Again, a visual inspection of our results (see Figure 6b) indicates that participants assigned to the condition involving the clear, eight-feature model had, on average, higher prediction errors than participants assigned to the other primary experimental conditions. Similar to our first experiment, this difference was significant ($t\,(594) = 4.78$, $p < 0.001$). Though again, we note that although these differences are statistically significant, they are not very large.

These findings motivated us to also investigate whether there were other differences between the condition involving the clear, eight-feature model and the other primary experimental conditions. For both the first and second experiment, we found that participants who were assigned to the condition involving the clear, eight-feature model were less good at simulating the model's predictions compared to participants assigned to the other primary experimental conditions ($t\,(994) = 7.96$, $p < 0.001$ for the first experiment, $t\,(594) = 7.23$, $p < 0.001$ for the second experiment; see Figures 3a and 4a) and that they were least likely to follow the model's predictions when it was beneficial for them to do so ($t\,(994) = 2.37$, $p = 0.018$ for the first experiment, $t\,(594) = 2.49$, $p = 0.012$ for the second experiment; see Figures 3b and 4b).

To summarize, the findings from these additional post-hoc analyses of the data from our first two experiments lend even more support to our explanation for the difference between the findings from our first two experiments and the findings from our third.

7 LIMITATIONS

One limitation of our work is that our experiments focused on one type of stakeholder (laypeople) using one type of model (linear regression) in one domain (real estate valuation). Future extensions to other types of stakeholders (e.g., data scientists, domain experts), other tasks (e.g., classification), other types of models (e.g, decision trees, rule lists, deep neural networks), and other domains (e.g., medical diagnosis, credit risk assessment, judicial sentencing and bail, hiring) may yield different findings.

In our first three experiments, we constrained the two-feature model and the eight-feature to make the same predictions. Although there are some domains where this is possible [48], there are of course others—such as computer vision and natural language processing—where more complex, deep models tend to outperform simpler ones. We did not experiment with such models because it would have created a confound, meaning that we would not have known whether any differences we observed were due to the presentation of the model, the model fidelity, or the very large number of features that complex, deep models typically use. Although our experiments did not involve complex, deep models, our main findings still have important implications for these domains: absent other reasons for using clear models, scientific evidence about what aids decision making the most should carry more weight than common intuition about interpretability. We also emphasize that, in our experiments, the conditions involving black-box models were designed to capture how people engage with models that could have arbitrarily complex internal structures, including, for instance, deep neural networks. Indeed, although readers of this paper know that we used linear regression models, participants in our experiments had no reason to believe that this was the case.

Even though our experiments were carefully designed and tightly controlled, we cannot rule out the possibility that other aspects of the models influenced our findings. For example, participants might have found the particular features used in the two-feature model (i.e., bathrooms and square feet) to be less intuitive than other possible combinations of two features (e.g., bedrooms and bathrooms) or even three features (e.g., bedrooms, bathrooms, and square feet). Also, participants who were shown the two-feature model had access to more information than the model—a scenario known in the decision-making literature as the "broken leg problem" [24, p. 151]. For this reason, they may have thought that the model was not relying on information that it should have. Perhaps if they were told that using the remaining six features would not improve the model's accuracy, they would have viewed the two-feature model differently. Conversely, though, it could have been the case that aspects of the eight-feature model led participants to question it. For instance, the negative coefficient for total rooms (which accounted for correlations between number of bedrooms and number of bathrooms) might have been confusing or mistakenly viewed as wrong, leading participants to follow the model's predictions less closely than they would have done otherwise.

Lastly, our experiments were run without process measures as dependent variables, which limited our ability to reflect on the cognitive and sensemaking processes that might have been at play. As one example, while we measure participants' ability to detect and correct for the model's sizable mistakes in terms of their deviation from the model on apartments with unusual configurations, we are unable to directly infer from these results whether participants understood why the model made these mistakes. Qualitative experiments (involving interviews, think aloud protocols, process-tracing measures, etc.), targeted at understanding *why* people behave in the ways they do, may be useful for investigating cognitive and sensemaking aspects of interpretability. On top of that, our experiments were short and one shot. Deeper insight into sensemaking could be gained not only by collecting process measures but by doing so longitudinally.

8 DISCUSSION AND CONCLUSION

Our experiments yielded some unexpected findings. First, we did not find a significant improvement in the extent to which participants followed the predictions of a clear model with few features compared to the predictions of a black-box model with more features. We also found that participants would have had lower prediction errors had they simply followed the model's predictions.

Furthermore, we found that using a clear model hampered participants' abilities to detect when the model had made a sizable mistake, seemingly due to information overload caused by the amount of detail in front of them. When we investigated an outlier focus message, intended to counter information overload, we found that this behavior disappeared. Several findings from our post-hoc

analyses are also consistent with the idea that too much information can be detrimental. In our first two experiments, in the condition in which participants were shown the *most* information—i.e., the condition involving the clear, eight-feature model—participants were *worst* at simulating the model's predictions, followed the model's predictions *less*, and made *less accurate* predictions of the apartments' selling prices compared to participants assigned to the other primary experimental conditions.

These findings suggest new ways to present models to people. When technically possible, it may be helpful to alert people when the data point in front of them may be an outlier. This could be achieved by training an auxiliary model to detect such data points. In addition, it may be prudent to ask people for their own predictions before seeing the model's predictions or even the model itself. Doing so could encourage people to inspect each data point carefully, making them more likely to notice any unusual feature values. Indeed, this idea is supported by recent research, which found that eliciting predictions and presenting feedback is beneficial for people's memory and comprehension of data points [52]. Lastly, despite the potential benefits of clear models, it may be detrimental to expose model internals by default, as doing so might cause people to experience information overload. Instead, model internals could be hidden until the person using the model requests to see them. Testing these suggestions empirically would be a natural direction for future research.

We emphasize that none of this is to say that the number of features or the transparency of the model should be ignored. Instead, our findings underscore the point that there are many possible goals when developing interpretable models, and that testing, not intuition, should be used to assess whether these goals have been met [65, 110].

Although we found that two factors commonly thought to make machine learning models more interpretable often have negligible effects on people's behavior and, in some cases, even have detrimental effects, there is still a long list of reasons why clear models with few features may be desirable. First, in some domains, transparency may play an important role in people's willingness to use a model on ethical grounds. For instance, if a model is used to aid judicial decision making, policy makers may demand transparency so as to be assured that the model does not rely on disallowed information, like race, or proxies for disallowed information. Second, access to model internals permits types of debugging or analyses that would otherwise be difficult. In fact, we leveraged this aspect of our linear regression models to generate some of the unusual apartment configurations used in our experiments, since we could easily see that the models would place an unreasonably high value on additional bathrooms when other feature values were held constant. Third, in scenarios where it is desirable to have a model that is easy to simulate, our findings suggest people can better simulate the predictions of clear models with few features. Fourth, although we did not investigate the field adoption of machine learning models, it might be the case that people are more likely to use simpler models than more complex ones because they find them more appealing [48]. Given that we did not find a large difference in participants' prediction errors between the primary experimental condition in our first two experiments, if people are more willing to use simpler models, there could be substantial benefits in terms of accuracy.

Given the widespread and increasing use of machine learning models, it is likely that people will make more and more decisions in collaboration with models. As this happens, it is also likely that there will be an increased demand for models that are interpretable. We hope that our work reinforces the importance of testing over intuition when developing interpretable models—i.e., what is or is not interpretable must be defined by people's behavior.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] Russell L Ackoff. 1967. Management misinformation systems. *Management Science* 14, 4 (1967), B–141–B–274. https://doi.org/10.1287/mnsc.14.4.B147

[3] Oscar Alvarado and Annika Waern. 2018. Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 286. http://doi.acm.org/10.1145/3173574.3173860

[4] Giuseppe Amatulli, Maria João Rodrigues, Marco Trombetti, and Raffaella Lovreglio. 2006. Assessing long-term fire risk at local scale by means of decision tree technique. *Journal of Geophysical Research: Biogeosciences* 111, G4 (2006).

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (2016).

[6] Thomas Åstebro and Samir Elhedhli. 2006. The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science* 52, 3 (2006), 395–409. https://doi.org/10.1287/mnsc.1050.0468

[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48.

[8] Max H Bazerman. 1985. Norms of distributive justice in interest arbitration. *ILR Review* 38, 4 (1985), 558–570.

[9] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: Human considerations in context-aware systems. *Human–Computer Interaction* 16, 2–4 (2001), 193–212.

[10] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD Cup and Workshop*.

[11] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377. http://doi.acm.org/10.1145/3173574.3173951

[12] L Breiman, JH Friedman, R Olshen, and CJ Stone. 1984. Classification and Regression Trees. (1984).

[13] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (2017), 30–44.

[14] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[15] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[16] Krista Casler, Lydia Bickel, and Elizabeth Hackett. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior* 29, 6 (2013), 2156–2160.

[17] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[18] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[19] Marvin M Chun. 2000. Contextual cueing of visual attention. *Trends in cognitive sciences* 4, 5 (2000), 170–178.

[20] Eric Colson. 2013. Using human and machine processing in recommendation systems. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[21] Alexander Coppock. 2019. Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods* 7, 3 (2019), 613–628.

[22] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. (2018).

[23] Robyn M Dawes. 1979. The robust beauty of improper linear models in decision making. *American psychologist* 34, 7 (1979), 571.

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

[25] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126.

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

[27] Jaap J Dijkstra. 1999. User agreement with incorrect expert system advice. *Behaviour & Information Technology* 18, 6 (1999), 399–411.

[28] Jaap J Dijkstra, Wim BG Liebrand, and Ellen Timminga. 1998. Persuasiveness of expert systems. *Behaviour & Information Technology* 17, 3 (1998), 155–163.
[29] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.
[30] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
[31] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1 (2002), 79–94.
[32] Dedre Gentner and Albert L. Stevens. 1983. *Mental Models*. Lawrence Erlbaum Associates.
[33] Vivian Giang. 2018. The Potential Hidden Bias In Automated Hiring Systems. (2018). Accessed at https://www.fastcompany.com/40566971/the-potential-hidden-bias-in-automated-hiring-systems/.
[34] Gerd Gigerenzer and Daniel G Goldstein. 1996. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review* 103, 4 (1996), 650.
[35] Gerd Gigerenzer and Peter M Todd. 1999. *Simple heuristics that make us smart*. Oxford University Press, USA.
[36] Francesca Gino and Don A. Moore. 2007. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making* 20, 1 (2007), 21–35.
[37] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI)*.
[38] Daniel G Goldstein and Gerd Gigerenzer. 2009. Fast and frugal forecasting. *International journal of forecasting* 25, 4 (2009), 760–772.
[39] Nitesh Goyal and Susan R Fussell. 2016. Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 288–302.
[40] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
[41] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment* 12, 1 (2000), 19.
[42] Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B Hamrick, and Patricia Chan. 2016. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods* 48, 3 (2016), 829–842.
[43] David J Hand and William E Henley. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160, 3 (1997), 523–541.
[44] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
[45] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (2011), 141–154.
[46] Jacob Jacoby. 1984. Perspectives on information overload. *Journal of consumer research* 10, 4 (1984), 432–435.
[47] Philip Johnson-Laird. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press.
[48] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2020. Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2020).
[49] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
[50] Kevin Lane Keller and Richard Staelin. 1987. Effects of quality and quantity of information on decision effectiveness. *Journal of consumer research* 14, 2 (1987), 200–213.
[51] Hyunjoong Kim, Wei-Yin Loh, Yu-Shan Shih, and Probal Chaudhuri. 2007. Visualizable and interpretable regression models with good prediction power. *IIE Transactions* 39, 6 (2007), 565–579.
[52] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1375–1386.
[53] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.
[54] Jon Kleinberg and Sendhil Mullainathan. 2019. Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 807–808.

[55] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[56] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

[57] Igor Kononenko. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine* 23, 1 (2001), 89–109.

[58] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. 2017. A Workflow for Visual Diagnostics of Binary Classifiers using Instance-Level Explanations. In *Proceedings of IEEE Conference and Visual Analytics Science and Technology*.

[59] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing*. 3–10.

[60] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 59–67.

[61] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2018. Human-in-the-Loop Interpretability Prior. In *Advances in Neural Information Processing Systems*.

[62] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1675–1684.

[63] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & Explorable Approximations of Black Box Models. In *FATML Workshop*.

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

[65] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[66] Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. 2018. Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 197–253.

[67] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*. ACM, 2119–2128. http://doi.acm.org/10.1145/1518701.1519023

[68] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[69] Jia Liu and Olivier Toubia. 2018. A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science* 37, 6 (2018), 855–1052. https://doi.org/10.1287/mksc.2018.1112

[70] Jennifer M. Logg. 2017. Theory of Machine: When Do People Rely on Algorithms? (2017). Harvard Business School NOM Unit Working Paper No. 17-086.

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

[72] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible Models for Classification and Regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (Beijing, China).

[73] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate Intelligible Models with Pairwise Interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (Chicago, IL, USA).

[74] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

[75] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[76] Paul E Meehl. 1990. Why summaries of research on psychological theories are often uninterpretable. *Psychological reports* 66, 1 (1990), 195–244.

[77] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2243–2251. http://doi.acm.org/10.1145/3269206.3272027

[78] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[79] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum. *arXiv preprint arXiv:1712.00547* (2017).
[80] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How search engines reinforce racism.* NYU Press.
[81] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).
[82] Don A. Norman. 1987. Some Observations on Mental Models. In *Human-Computer Interaction: A Multidisciplinary Approach*, R. M. Baecker and W. A. S. Buxton (Eds.). Morgan Kaufmann Publishers Inc., 241–244.
[83] Dilek Önkal, Paul Goodwin, Mary Thomson, and Sinan Gönül. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22 (2009), 390–409.
[84] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science* 23, 3 (2014), 184–188.
[85] Peter Pirolli and Daniel M Russell. 2011. Introduction to this special issue on sensemaking.
[86] Marianne Promberger and Jonathan Baron. 2006. Do patients trust computers? *Journal of Behavioral Decision Making* 19, 5 (2006), 455–468.
[87] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 103. http://doi.acm.org/10.1145/3173574.3173677
[88] Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*. ACM, 173–182. http://doi.acm.org/10.1145/2702123.2702174
[89] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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 (KDD)*.
[90] Christian Rudder. 2014. *Dataclysm: Love, Sex, Race, and Identity–What Our Online Lives Tell Us about Our Offline Selves*. Crown.
[91] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[92] Cynthia Rudin and Berk Ustun. 2018. Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. *Applied Analytics* 48, 5 (2018), 449–466. https://doi.org/10.1287/inte.2018.0957
[93] Chris Russell. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 20–28.
[94] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.
[95] Paul JH Schoemaker and C Carter Waid. 1982. An experimental comparison of different approaches to determining weights in additive utility models. *Management Science* 28, 2 (1982), 182–196. https://doi.org/10.1287/mnsc.28.2.182
[96] Kathryn Sharpe Wessling, Joel Huber, and Oded Netzer. 2017. MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research* 44, 1 (2017), 211–230.
[97] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662.
[98] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 303–310.
[99] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.
[100] Richard Tomsett, David Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems. In *2018 Workshop on Human Interpretability in Machine Learning*.
[101] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.
[102] Berk Ustun and Cynthia Rudin. 2016. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning Journal* 102, 3 (2016), 349–391.
[103] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 10–19.
[104] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors*

in Computing Systems. 1–13.

[105] Vanya MCA Van Belle, Ben Van Calster, Dirk Timmerman, Tom Bourne, Cecilia Bottomley, Lil Valentin, Patrick Neven, Sabine Van Huffel, Johan AK Suykens, and Stephen Boyd. 2012. A mathematical model for interpretable clinical decision support with applications in gynecology. *PloS one* 7, 3 (2012).

[106] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[107] Kim J Vicente and Gerard L Torenvliet. 2000. The Earth is spherical($p < 0.05$): alternative methods of statistical inference. *Theoretical Issues in Ergonomics Science* 1, 3 (2000), 248–271.

[108] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[109] Sara Wachter-Boettcher. 2017. Why You Can't Trust AI to Make Unbiased Hiring Decisions. (2017). Accessed at http://time.com/4993431/ai-recruiting-tools-do-not-/.

[110] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[111] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial Intelligence and Statistics*. 1013–1022.

[112] Martin Wattenberg, Fernanda Viégas, and Moritz Hardt. 2016. Attacking discrimination with smarter machine learning. (2016). Accessed at https://research.google.com/bigpicture/attacking-discrimination-in-ml/.

[113] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

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

APPENDICES

Appendix A SCENARIOS WHERE USERS HAVE ACCESS TO MORE INFORMATION THAN MODELS

Domain	Information the model uses	Side information a user has that the model does not	Citation
Malignancy risk in mammography	Age tumor density, 5 binary variables describing tumor shape	Full mammogram image, full medical records, clinical interview (habits, family history, etc), plus 10 binary variables not in interpretable model	[111]
Hospital readmission risk	7 binary features (bed sores, mood problems)	Full medical records, plus 23 other features not in interpretable model ("chronic pain", "feels unsafe", etc.)	[111]
Wildfire risk	29 binary features, mostly covering terrain type and temperature	Experience of past fires. local knowledge: history, hazardous industries, previous arson, etc. In addition, information not used by model: Continuous values of 9 variables, Any value of 4 variables, 8 dichotomous variables	[4]
Bail decisions	Statistical information available to judges at time of inquiry except disallowed ones	Information that is not used by the model because it is not allowed: Race, ethnicity, gender. Physical appearance of the defendant (e.g., "tattoos"), answers to questions, apparent remorse, etc..	[53]
Pre-trial release decisions	7 binary variables covering age and past failures to appear	Physical appearance of the defendant, answers to questions, apparent remorse, etc. Information not used by the interpretable model: 49 features describing the charges, 13 characteristics of the defendant.	[48]
Success of early-stage ventures	21 trinary features of companies	Industry knowledge and experience, subjective assessments that lead to trinary scores, and 16 trinary features not used by the interpretable model	[6]
Housing price prediction	3 features: number of rooms, % lower-income citizens, student-teacher ratio	Physical walkthrough of property, neighborhood knowledge. Continuous values of the 3 features used by the model, 10 continuous variables not used by the model.	[51]

Domain	Information the model uses	Side information a user has that the model does not	Citation
Baseball player salary prediction	3 variables: number of years in major leagues, career hits, hits in previous year	Personal experience watching players. 19 variables not included in the interpretable model.	[51]
Sleep apnea screening	5 binary features	Medical records, patient interview, 28 binary features not included in the interpretable model.	[102]
Classification of high- and low-risk heart attack patients	3 binary features	Patient interview. Continuous measures of these 3 features. 16 features collected at intake but not included in the interpretable model.	[35]; [12]
Prediction of non-viable pregnancies	6 features, each cut into 2 to 5 bins: maternal age, bleeding score, gestational age, gestational sac diameter, yolk sac diameter, fetal heart beat.	Continuous values on all variables. Any other information in medical records, patient interview, etc.	[105]

Table 1. Examples of decision aids whose users have access to more information than the models do.

Appendix B INSTRUCTIONS FROM THE CLEAR-2 CONDITION IN EXPERIMENT 1

The following instructions were shown to participants assigned to the CLEAR-2 condition in our first experiment on Mechanical Turk. The instructions for other conditions and experiments were adapted from these instructions with minimal changes.

Instructions

!! IMPORTANT !! Your session will expire in 60 minutes. Please make sure to complete the HIT in 60 minutes!

- You are here to predict **New York City apartment prices in the Upper West Side** with the help of a model.
- There will be a training phase and a testing phase:
  - In the training phase, you will see examples of apartments along with what the model predicted they sold for and the actual price they sold for.
  - In the testing phase, you will see new apartments and make your own prediction about what the model will predict and what the actual price is.

Next ➜

Instructions

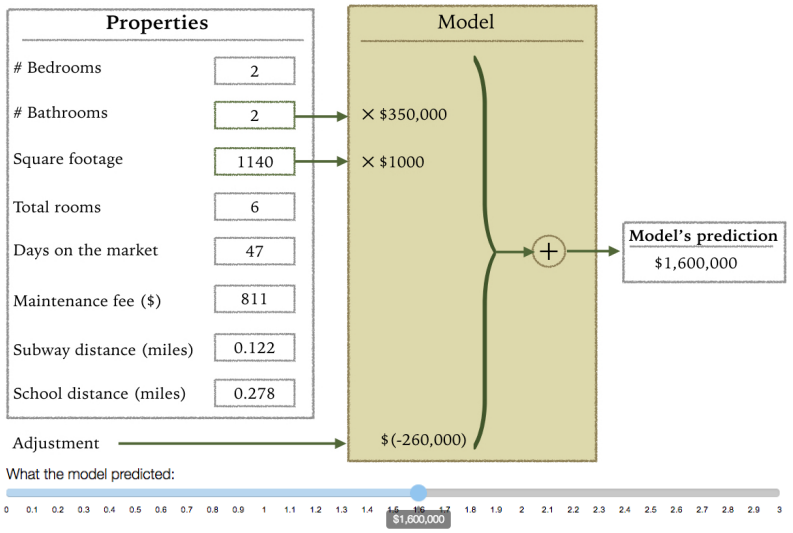
- You will see these properties for each apartment:

Properties	
# Bedrooms	2
# Bathrooms	2
Square footage	1140
Total rooms	6
Days on the market	47
Maintenance fee ($)	811
Subway distance (miles)	0.122
School distance (miles)	0.278

← Previous Next →

Instructions

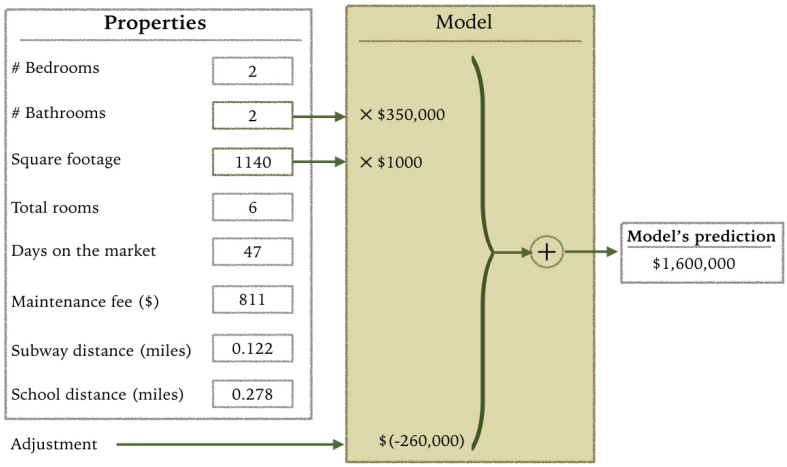
- A model predicts apartment prices. We will explain how this model works in the next page.
  - This model uses # Bathrooms and Square footage of the apartment to make its prediction.
  - The graph at the bottom shows this price visually.



← Previous Next →

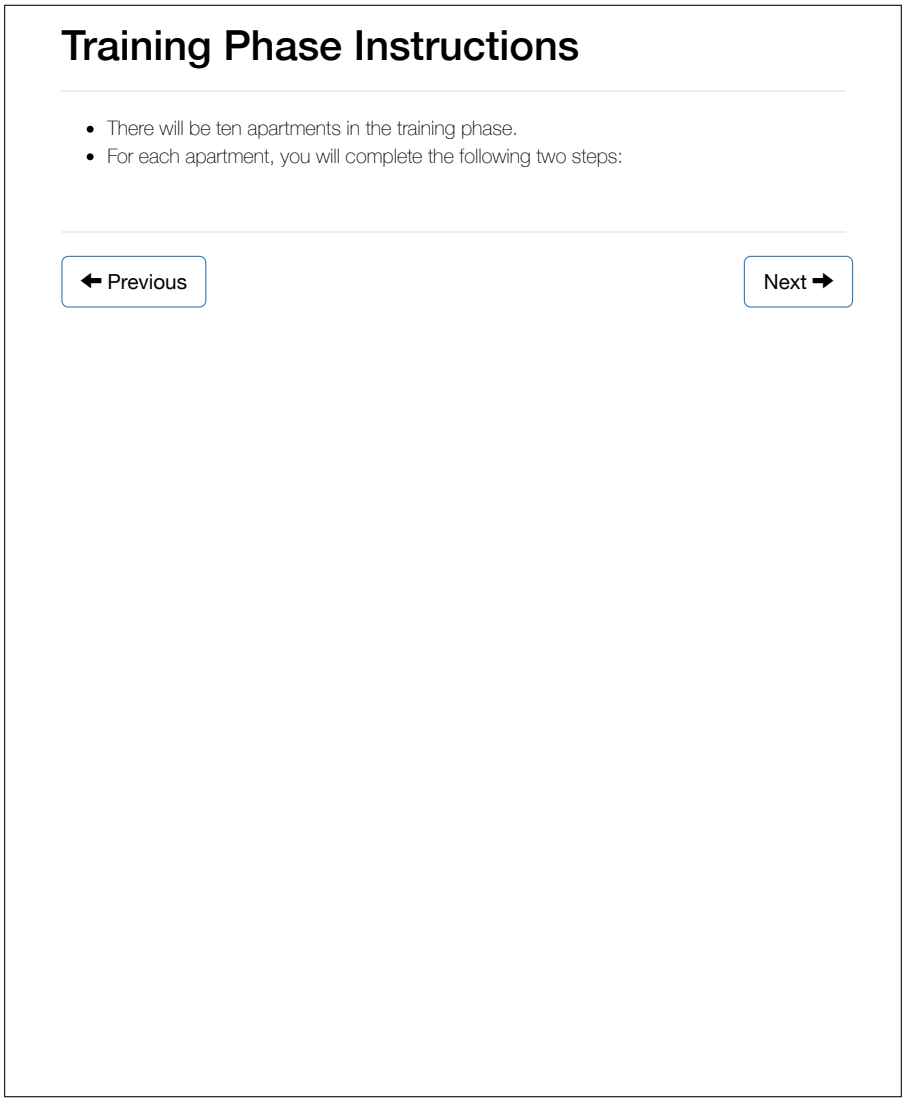
Instructions

- Here is how the model has made its prediction:
- Each bathroom is worth $350,000. Therefore, $350,000 is multiplied by the number of bathrooms and added to the price. This is repeated for Square footage. Finally, **the adjustment factor of $260,000 is subtracted** and a price is predicted.



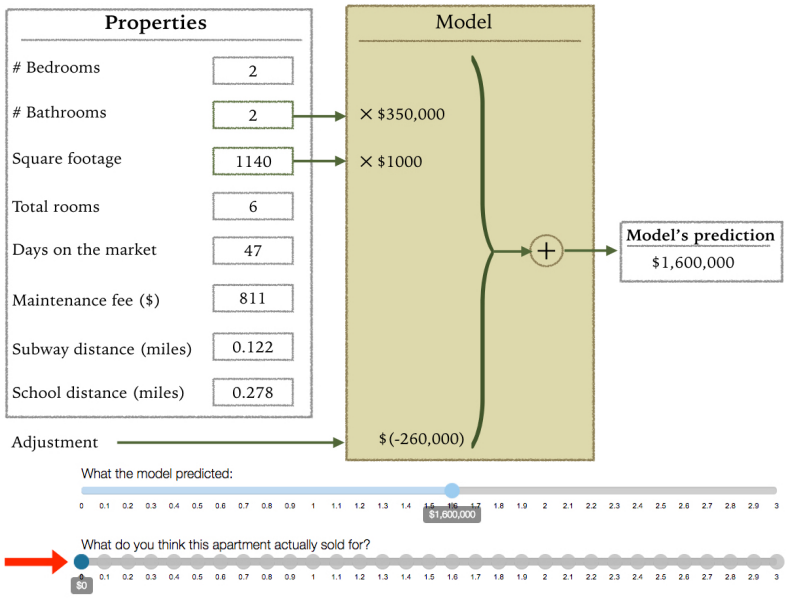
$$[2 \times 350{,}000] + [1140 \times 1000] + [-260{,}000]$$
$$700{,}000 + 1{,}140{,}000 + [-260{,}000]$$
$$\approx 1{,}600{,}000$$

← Previous Next →

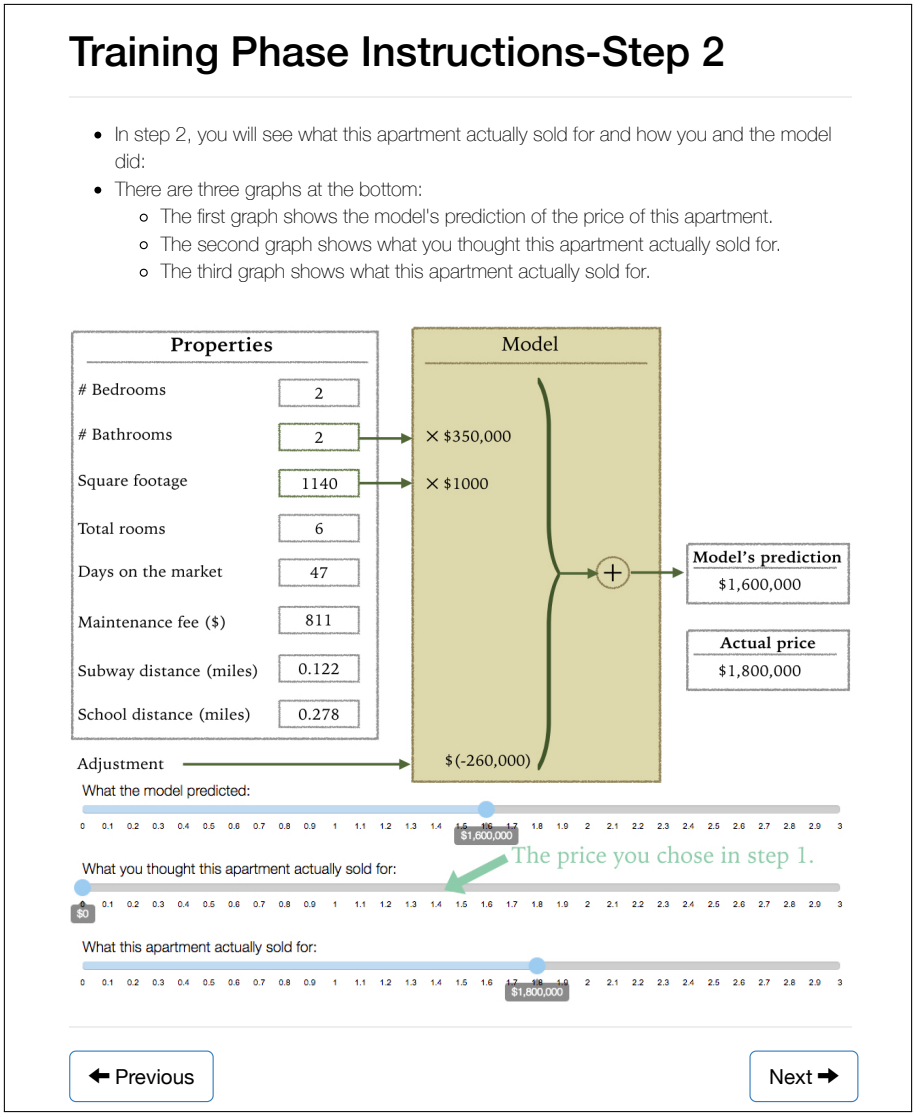

Training Phase Instructions
There will be ten apartments in the training phase.
For each apartment, you will complete the following two steps:
Previous
Next

Training Phase Instructions-Step 1

- In step 1, given the model's prediction, you will state what you think the apartment actually sold for:



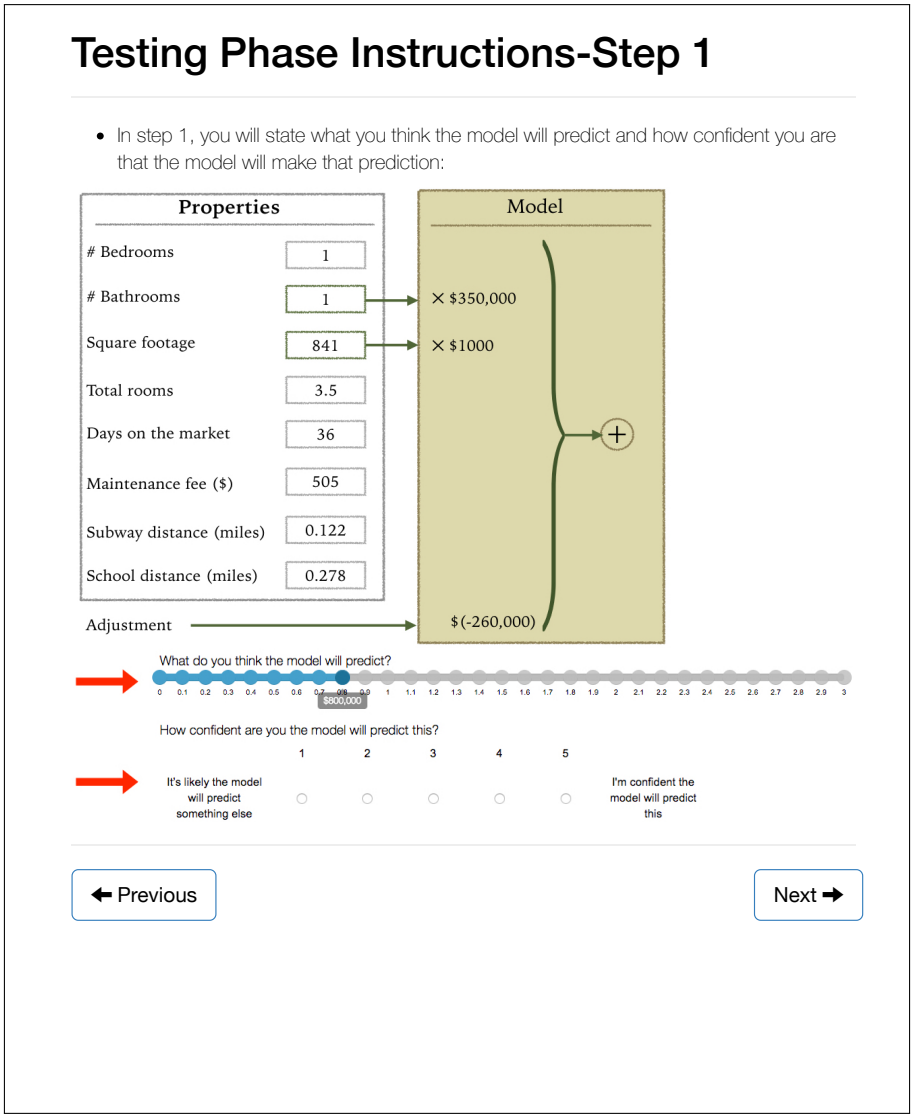
← Previous Next →

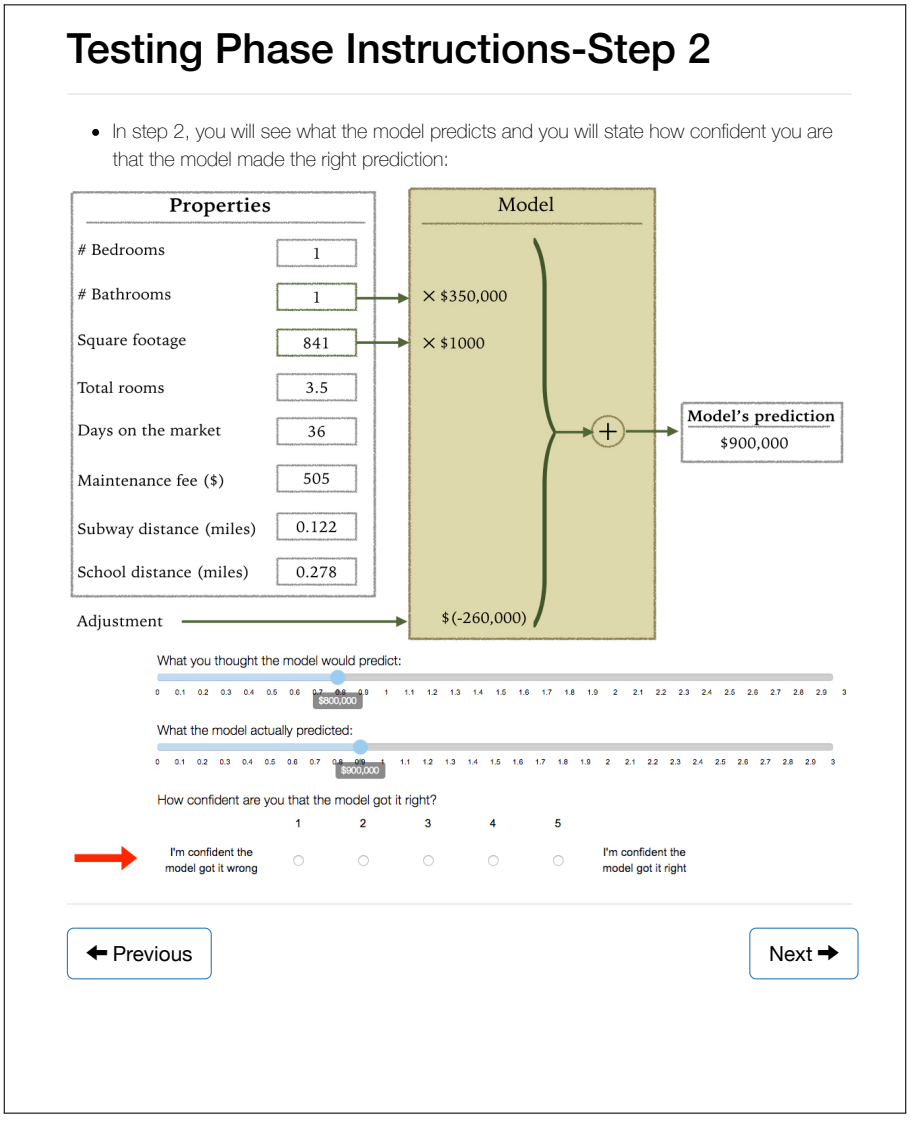

Training Phase Instructions-Step 2
In step 2, you will see what this apartment actually sold for and how you and the model did:
There are three graphs at the bottom:
The first graph shows the model's prediction of the price of this apartment.
The second graph shows what you thought this apartment actually sold for.
The third graph shows what this apartment actually sold for.
Properties
Bedrooms 2
Bathrooms 2
Square footage 1140
Total rooms 6
Days on the market 47
Maintenance fee ($) 811
Subway distance (miles) 0.122
School distance (miles) 0.278
Model
× $350,000
× $1000
$(-260,000)
Adjustment
Model's prediction
$1,600,000
Actual price
$1,800,000
What the model predicted:
$1,600,000
What you thought this apartment actually sold for:
The price you chose in step 1.
$0
What this apartment actually sold for:
$1,800,000
Previous
Next

Instructions

- Once you have reviewed all ten apartments in the training phase, you will move to the testing phase.
- In the testing phase, you will see twelve new apartments and you will guess the price each was sold for.
- **NOTE**: You will not see the actual prices for these apartments in the testing phase. Once you are done, you will see how you did overall.
- For each apartment, you will complete the following three steps:

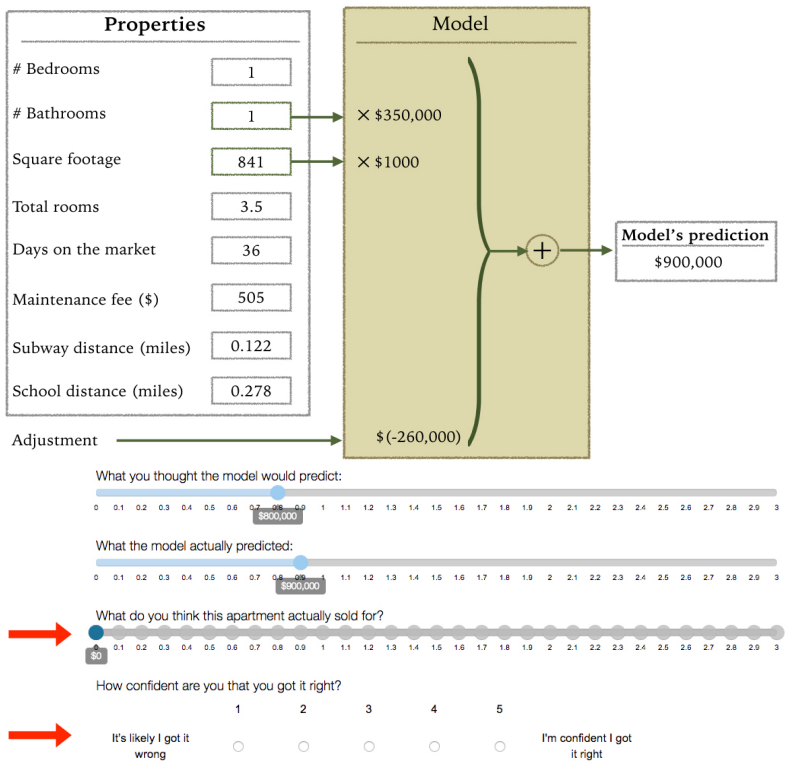
← Previous Next →


Testing Phase Instructions-Step 1
In step 1, you will state what you think the model will predict and how confident you are that the model will make that prediction:
Properties
Bedrooms 1
Bathrooms 1
Square footage 841
Total rooms 3.5
Days on the market 36
Maintenance fee ($) 505
Subway distance (miles) 0.122
School distance (miles) 0.278
Model
× $350,000
× $1000
Adjustment
$(-260,000)
What do you think the model will predict?
How confident are you the model will predict this?
1 2 3 4 5
It's likely the model will predict something else
I'm confident the model will predict this
← Previous
Next →


Testing Phase Instructions-Step 2
In step 2, you will see what the model predicts and you will state how confident you are that the model made the right prediction:
Properties
Bedrooms 1
Bathrooms 1
Square footage 841
Total rooms 3.5
Days on the market 36
Maintenance fee ($) 505
Subway distance (miles) 0.122
School distance (miles) 0.278
Adjustment
Model
× $350,000
× $1000
$(-260,000)
Model's prediction
$900,000
What you thought the model would predict:
What the model actually predicted:
How confident are you that the model got it right?
1 2 3 4 5
I'm confident the model got it wrong
I'm confident the model got it right
Previous
Next

Testing Phase Instructions-Step 3

- In step 3, given the model's prediction, you will state what you think this apartment actually sold for and your confidence:



← Previous Next →

Instructions

- Once you are done with twelve apartments in the testing phase, you will see your results and how you did overall.
- You are now done with all the instructions. Thanks for participating in this experiment!

You must provide correct answer to the following question to proceed:

Each apartment has the following 8 properties:

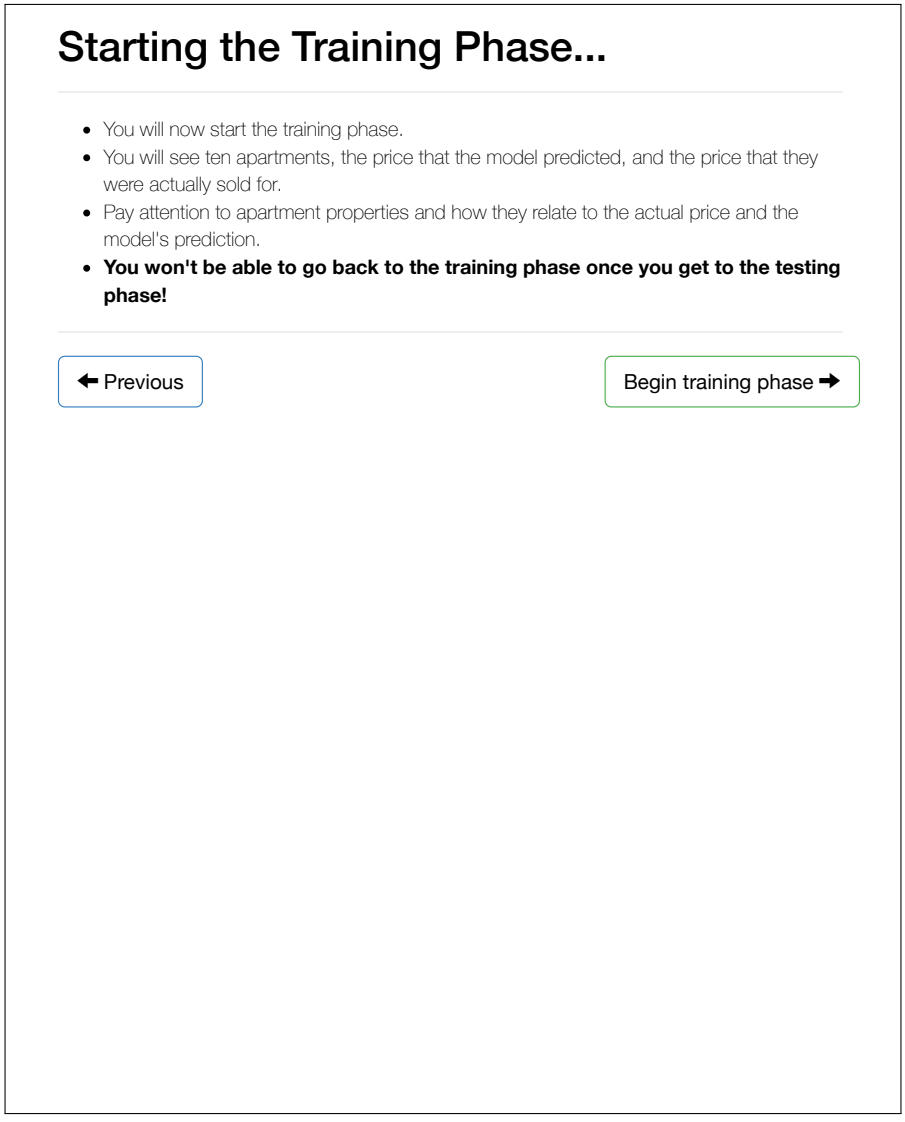
Bedrooms, # Bathrooms, Square footage, Total rooms, Days on the market, Maintenance fee, Subway distance, and School distance

How many of these apartment properties does the model use to make its prediction?

- ◯ 1 property
- ◯ 2 properties
- ◯ 3 properties
- ◯ 4 properties
- ◯ 5 properties
- ◯ 6 properties
- ◯ 7 properties
- ◯ 8 properties

Submit

← Previous Next →


Starting the Training Phase...
You will now start the training phase.
You will see ten apartments, the price that the model predicted, and the price that they were actually sold for.
Pay attention to apartment properties and how they relate to the actual price and the model's prediction.
You won't be able to go back to the training phase once you get to the testing phase!
Previous
Begin training phase

Appendix C APARTMENT SELECTION DETAILS

We used the following procedure to construct a set of ten apartments which are representative in terms of the models' prediction errors ($m - a$). First we selected all apartments for which the rounded predictions of the two- and eight-feature models agreed. Then we randomly sampled 5,000 sets of ten such apartments, computed the errors the model made on each apartment, and sorted them within each set to obtain the largest error, second largest error, and so on. We then computed the average largest error across all 5,000 sets and rounded it to the nearest $100K. We repeated this for the second through tenth largest errors. This resulted in the following ten average error values: -$500K, -$300K, -$200K, -$200K, -$100K, $0, $0, $100K, $100K, $300K.

For each of these ten error values, we randomly selected two apartments (for which the difference between the rounded model prediction and the rounded actual price matched the error value) and we randomly assigned one to the training and one to the testing phase. To ensure participants would see a good variety of apartment configurations—defined as the combination of number of bedrooms and number of bathrooms—this process was repeated until neither the training nor the testing set contained more than three apartments with the same configuration. Tables 2 and 4 show the configurations of apartments that were used during the training and testing phase of each of our experiments, respectively. Tables 3 and 5 show the predictions and errors of the two- and eight-feature models on each of the apartments.

Apartment ID	Apartment configurations							
	Bedrooms	Bathrooms	Square footage	Total rooms	Days on the market	Maintenance fee	Distance from the subway (miles)	Distance from a school (miles)
1	1	1	750	3	51	947	0.179	0.104
2	1	1	550	3	90	409	0.122	0.278
3	2	1	800	4	36	1160	0.218	0.365
4	2	1	850	4	30	1720	0.105	0.153
5	1	1	550	3	135	442	0.231	0.124
6	0	1	540	2.5	72	332	0.064	0.271
7	3	2	1990	6	213	1280	0.183	0.329
8	2	1	1150	4	37	1500	0.129	0.351
9	0	1	540	2.5	59	331	0.064	0.271
10	2	2	1300	5	39	1110	0.110	0.250

Table 2. Configuration of the apartments used in experiments 1, 2, and 3 during the training phase. In Experiment 4, apartments 4, 5, 6, 8, and 10 were used.

Apartment ID	two-feature model			eight-feature model		
	prediction	error	error fraction	prediction	error	error fraction
1	840,000	-9,000	0.011	768,930	-80,070	0.094
2	640,000	-10,000	0.015	632,010	-17,990	0.028
3	890,000	241,000	0.371	893,500	244,500	0.377
4	940,000	115,000	0.139	850,600	25,600	0.031
5	640,000	-184,000	0.223	614,880	-209,120	0.254
6	630,000	175,000	0.385	550,080	95,080	0.209
7	2,430,000	-470,000	0.162	2,417,800	-482,200	0.166
8	1,240,000	90,000	0.078	1,195,600	45,600	0.04
9	630,000	-165,000	0.208	552,790	-242,210	0.305
10	1,740,000	-260,000	0.13	1,701,100	-298,900	0.149

Table 3. Prediction, prediction error (i.e., $m - a$), and prediction error fraction (i.e., $(m - a)/a$) of the two- and eight-feature models on the apartments used in the training phase of our experiments.

Apartment ID	Apartment configurations							
	Bedrooms	Bathrooms	Square footage	Total rooms	Days on the market	Maintenance fee	Distance from the subway (miles)	Distance from a school (miles)
1	1	1	925	3	80	954	0.173	0.312
2	2	1	1080	5	39	846	0.207	0.212
3	3	2	1530	5	15	1550	0.226	0.251
4	2	2	1140	4.5	93	863	0.122	0.278
5	1	1	540	3	11	437	0.202	0.199
6	0	1	540	2.5	74	341	0.122	0.278
7	2	1	1240	4.5	32	1370	0.081	0.262
8	2	2	1240	4.5	14	906	0.178	0.225
9	2	1	1250	5	23	1480	0.089	0.281
10	1	1	532	2.5	20	388	0.122	0.278
11	1	2	750	3	225	825	0.159	0.144
12	1	3	726	4	17	444	0.121	0.101
13	1	1	788	3.5	51	473	0.122	0.278
14	1	3	350	4	13	430	0.221	0.131

Table 4. Configuration of the apartments used in our experiments during the testing phase. Apartments 1–12 were used in experiments 1, 2, and 3. In Experiment 4, apartments 1, 6, 8, 9, 10, 12 ("Apartment 6" in Experiment 4), 13 ("Apartment 7" in Experiment 4), and 14 ("Apartment 8" in Experiment 4) were used. Apartments 12 and 14 were synthetically generated.

Apartment ID	two-feature model			eight-feature model		
	prediction	error	error fraction	prediction	error	error fraction
1	1,015,000	90,000	0.097	957,560	32,560	0.035
2	1,170,000	-80,000	0.064	1,166,040	-83,960	0.067
3	1,970,000	-560,000	0.221	1,989,200	-540,800	0.214
4	1,580,000	-170,000	0.097	1,573,970	-176,030	0.100
5	630,000	74,000	0.133	634,830	78,830	0.142
6	630,000	-145,000	0.187	555,190	-219,810	0.284
7	1,330,000	331,000	0.331	1,274,700	275,700	0.276
8	1,680,000	-20,000	0.012	1,685,340	-14,660	0.009
9	1,340,000	-210,000	0.135	1,264,600	-285,400	0.184
10	622,000	97,000	0.185	642,820	117,820	0.224
11	1,190,000	541,000	0.834	1,099,550	450,550	0.694
12	1,516,000	—	—	1,475,960	—	—
13	878,000	-292,000	0.25	858,270	-311,730	0.266
14	1,140,000	—	—	1,115,300	—	—

Table 5. Prediction, prediction error (i.e., $m - a$), and prediction error fraction (i.e., $(m - a)/a$) of the two- and eight-feature models on the apartments used in the testing phase of our experiments.

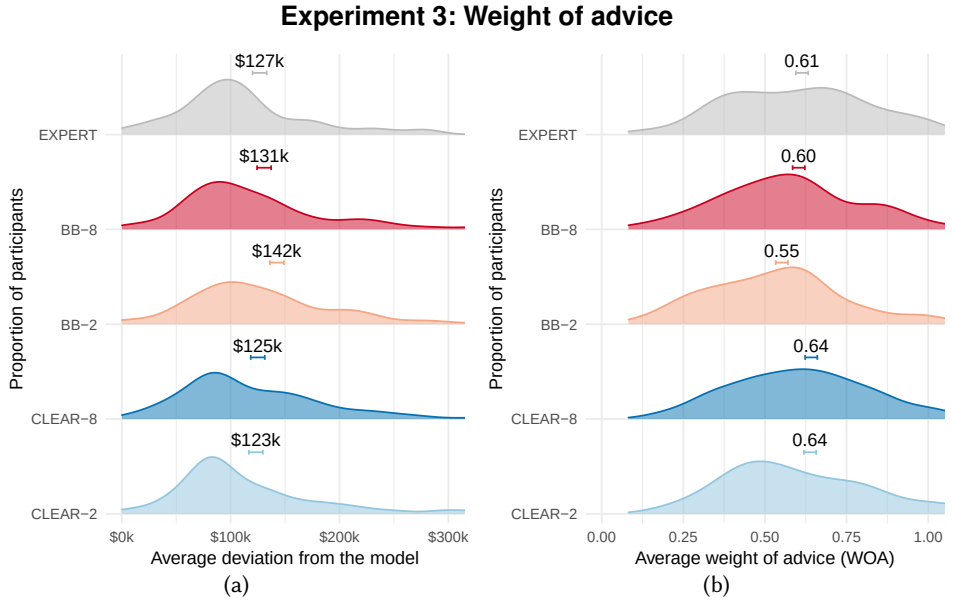


Fig. 10. Results from Experiment 3: density plots for (a) mean deviation of participants' predictions from the model's prediction and (b) mean weight of advice. Numbers in each subplot indicate average values over all participants in the corresponding condition and error bars indicate one standard error.

Appendix D EXPERIMENT 3 HYPOTHESES AND FINDINGS

We pre-registered four hypotheses:[10]

- H7. **Deviation.** Participants' predictions will deviate less from the predictions of a clear model with a small number of features than the predictions of a black-box model with a large number of features.
- H8. **Weight of advice.** Weight of advice will be higher for participants who see a clear model with a small number of features than for those who see a black-box model with a large number of features.
- H9. **Humans vs. machines.** Participants' deviation and weight of advice measures will differ depending on whether the predictions come from a black-box model with a large number of features or a human expert.
- H10. **Detection of mistakes.** Participants in different conditions will exhibit varying abilities to correct the model's inaccurate predictions on unusual examples.

The first two hypotheses are variations on H2 from our first experiment, while the last hypothesis is identical to H3.

D.1 Results

H7. **Deviation.** In line with the findings from the first two experiments, there was no significant difference in participants' deviation from the model between CLEAR-2 and BB-8 ($t\,(798) = -0.87$, $p = 0.384$, see Figure 10a).

H8. **Weight of advice.** Weight of advice is not well defined when a participant's initial prediction matches the model's prediction (i.e., $u_1 = m$). For each condition, we therefore calculated the mean

[10]Pre-registered hypotheses this experiment are available at https://aspredicted.org/795du.pdf.

weight of advice over all participant–apartment pairs for which the participant's initial prediction did not match the model's prediction, which can be viewed as calculating the mean conditional on there being a difference between the participant's and the model's predictions. Between conditions, we found no significant difference in the fraction of times that participants' initial predictions matched the model's predictions. In line with the findings for deviation in the first two experiments, there was no significant difference in participants' weight of advice between the CLEAR-2 and BB-8 conditions ($t\,(819) = 1.27$, $p = 0.205$, see Figure 10b).

H9. **Humans vs. machines.** The hypothesis that people would deviate less from machine predictions was not supported as there was not a significant difference in participants' deviation from the model ($t\,(994) = 0.45$, $p = 0.655$) or in their weight of advice ($t\,(1005) = -0.38$, $p = 0.704$) between the BB-8 and EXPERT conditions. We expect that the difference between our results and those in [70, 71] is due to participants getting more experience with the model (or expert) and its predictions over the course of twelve apartments in our experiment.

H10. **Detection of mistakes.** Participants in the clear conditions were no less able to correct inaccurate predictions ($t\,(798) = -0.96$, $p = 0.337$ and $t\,(798) = -0.19$, $p = 0.847$ for the contrast of CLEAR-2 and CLEAR-8 with BB-2 and BB-8 for apartments 11 and 12, respectively). We investigate this further in Experiment 4 (Section 6).

Appendix E FULL DISTRIBUTIONS OF PARTICIPANTS' PREDICTIONS

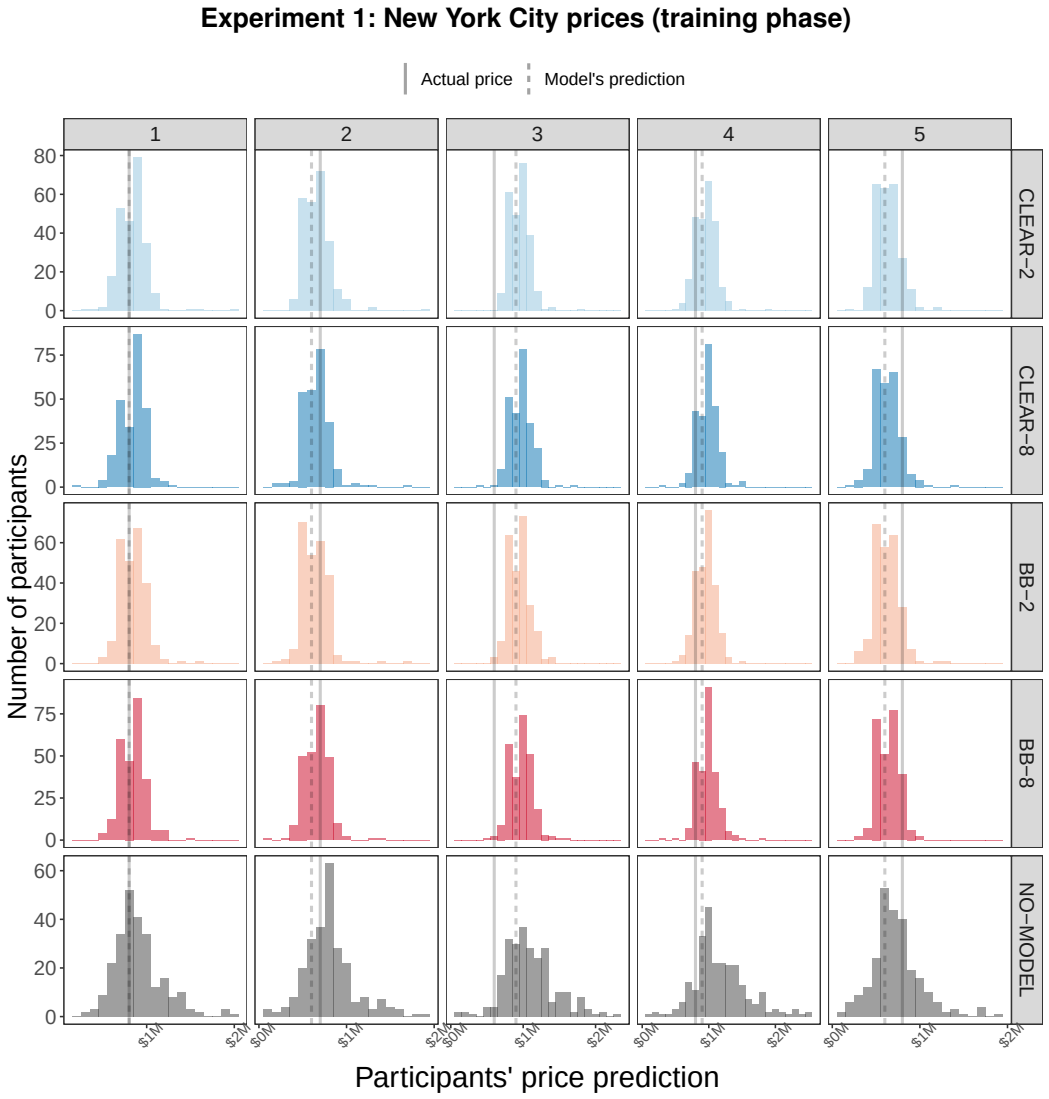


Fig. 11. Distribution of participants' predictions of prices of apartments 1–5 in the training phase in Experiment 1.

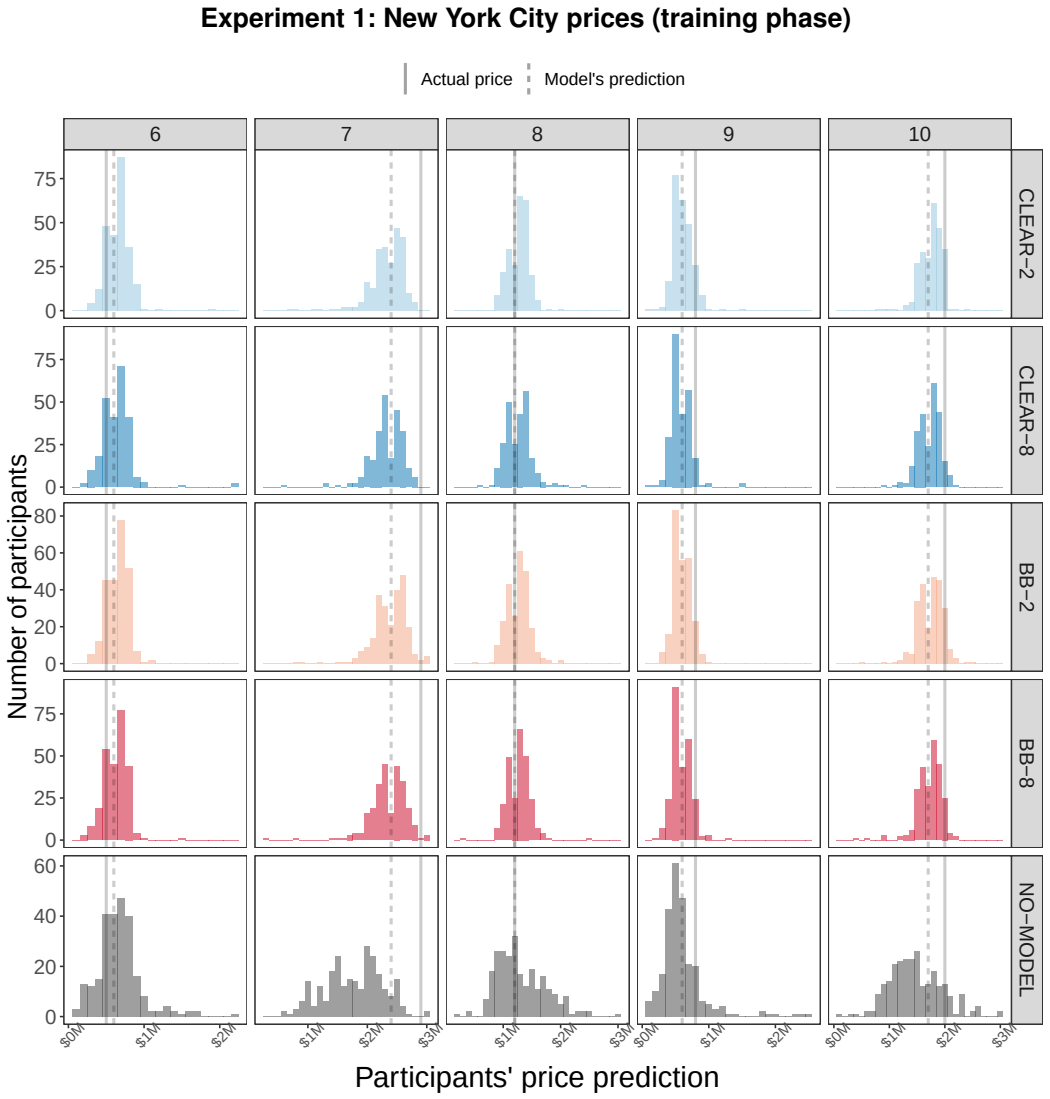


Fig. 12. Distribution of participants' predictions of prices of apartments 6–10 in the training phase in Experiment 1.

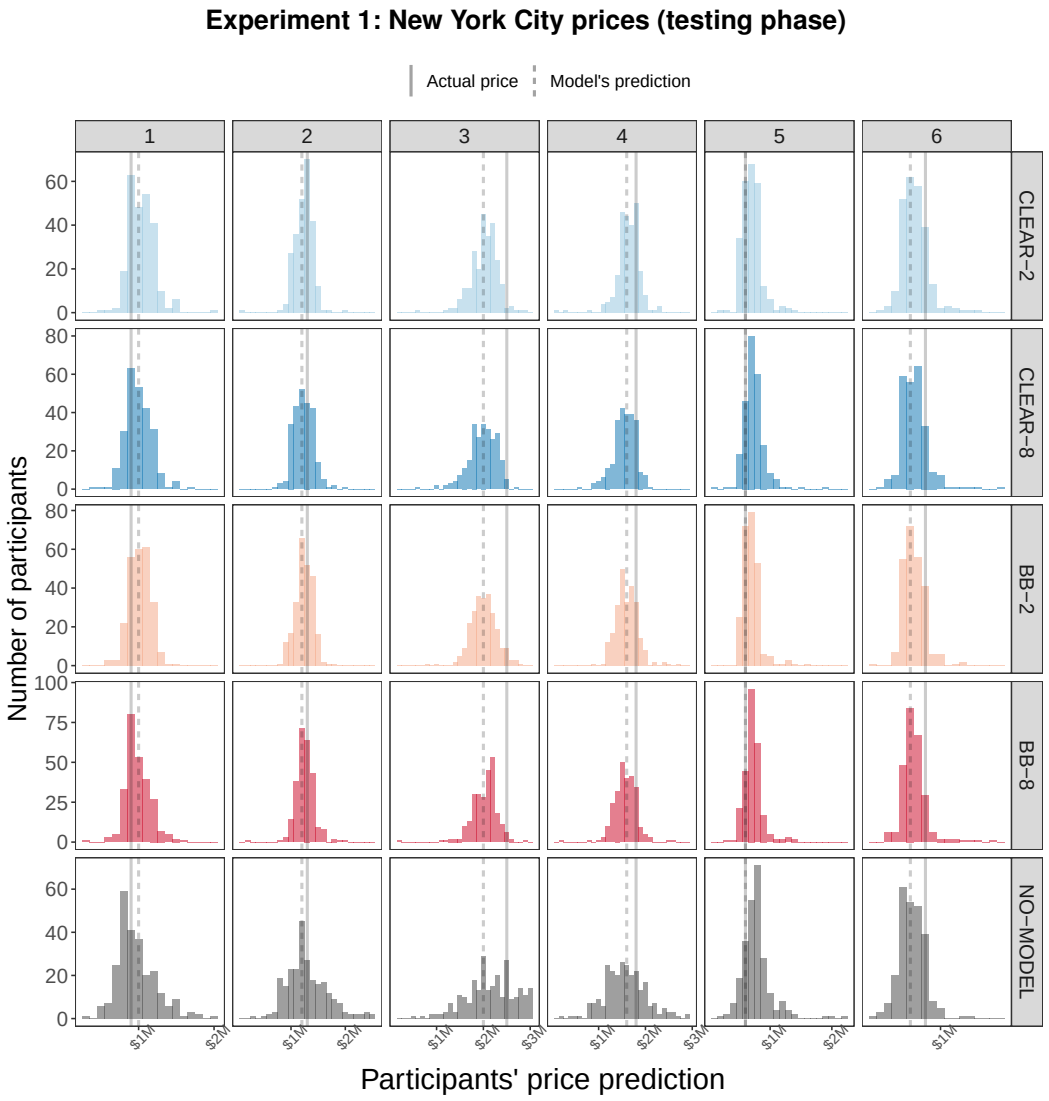


Fig. 13. Distribution of participants' predictions of prices of apartments 1–6 in the testing phase in Experiment 1.

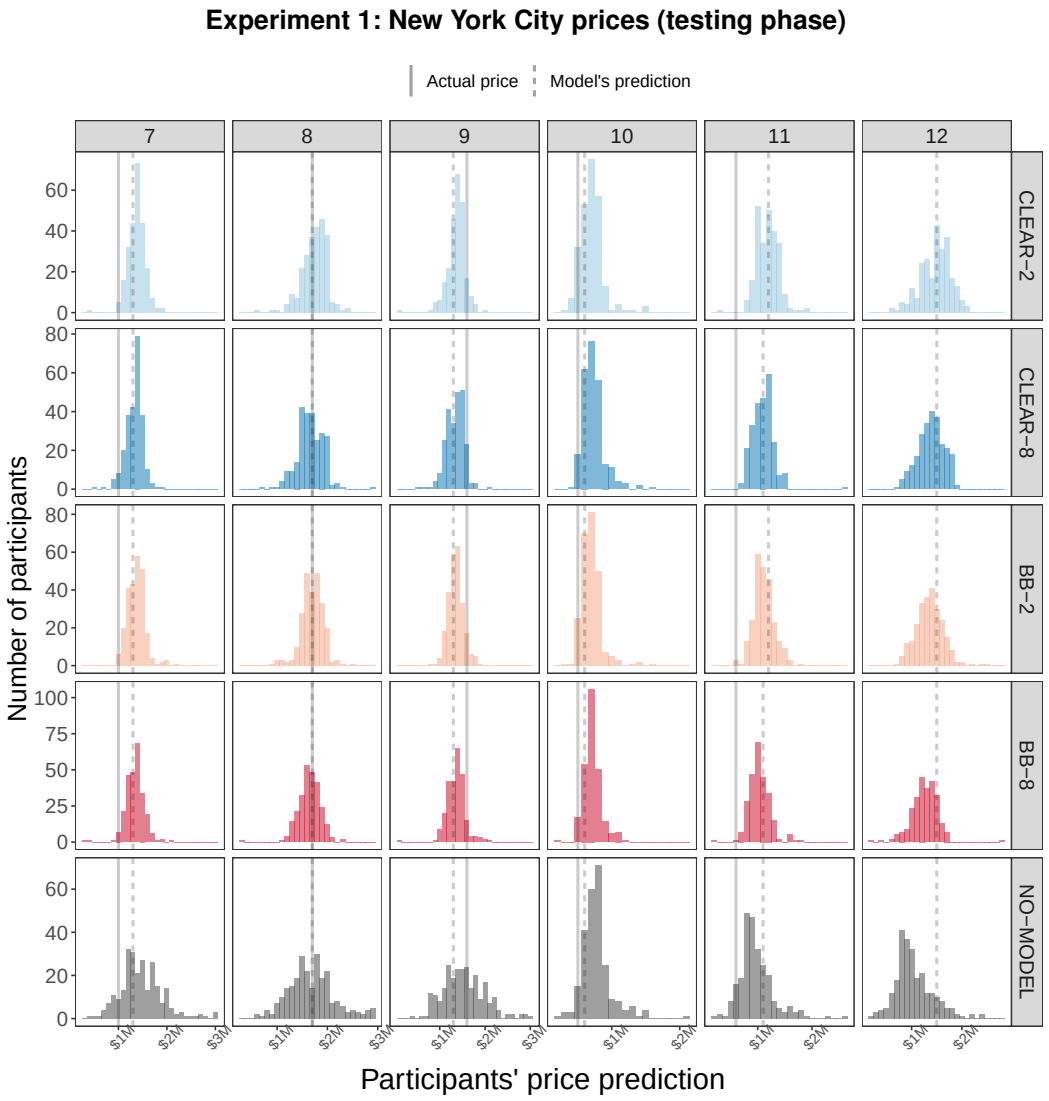


Fig. 14. Distribution of participants' predictions of prices of apartments 7–12 in the testing phase in Experiment 1.

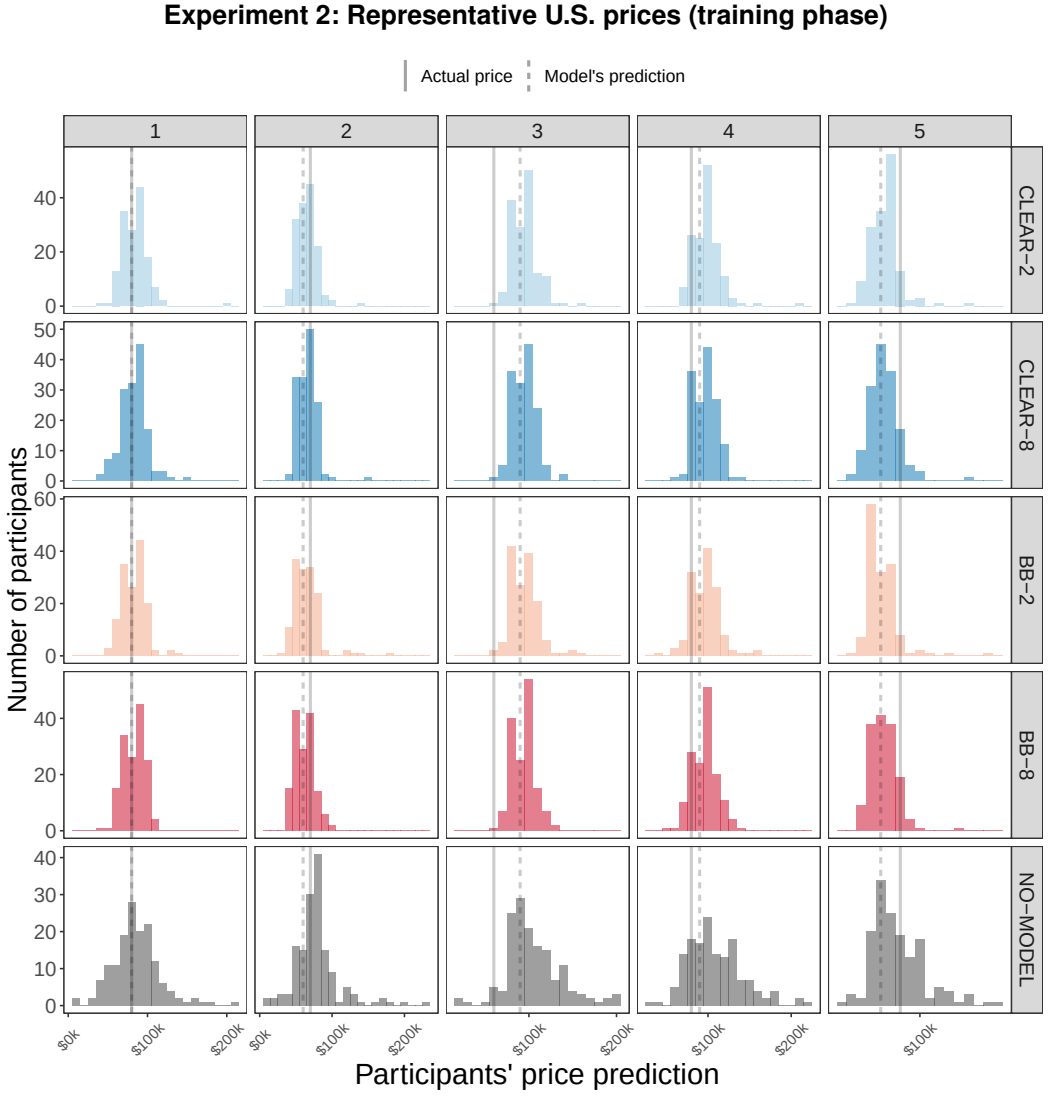


Fig. 15. Distribution of participants' predictions of prices of apartments 1–5 in the training phase in Experiment 2.

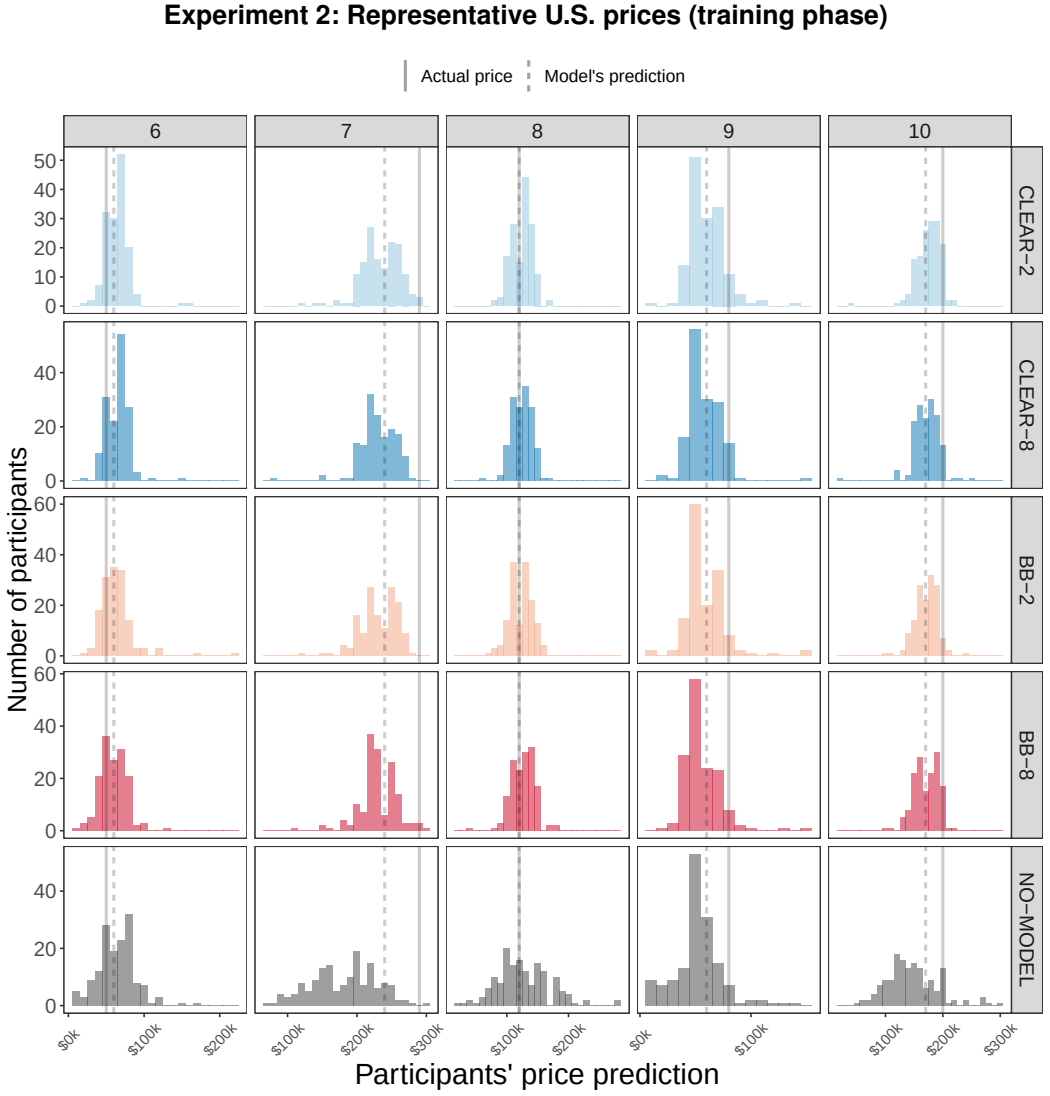


Fig. 16. Distribution of participants' predictions of prices of apartments 6–10 in the training phase in Experiment 2.

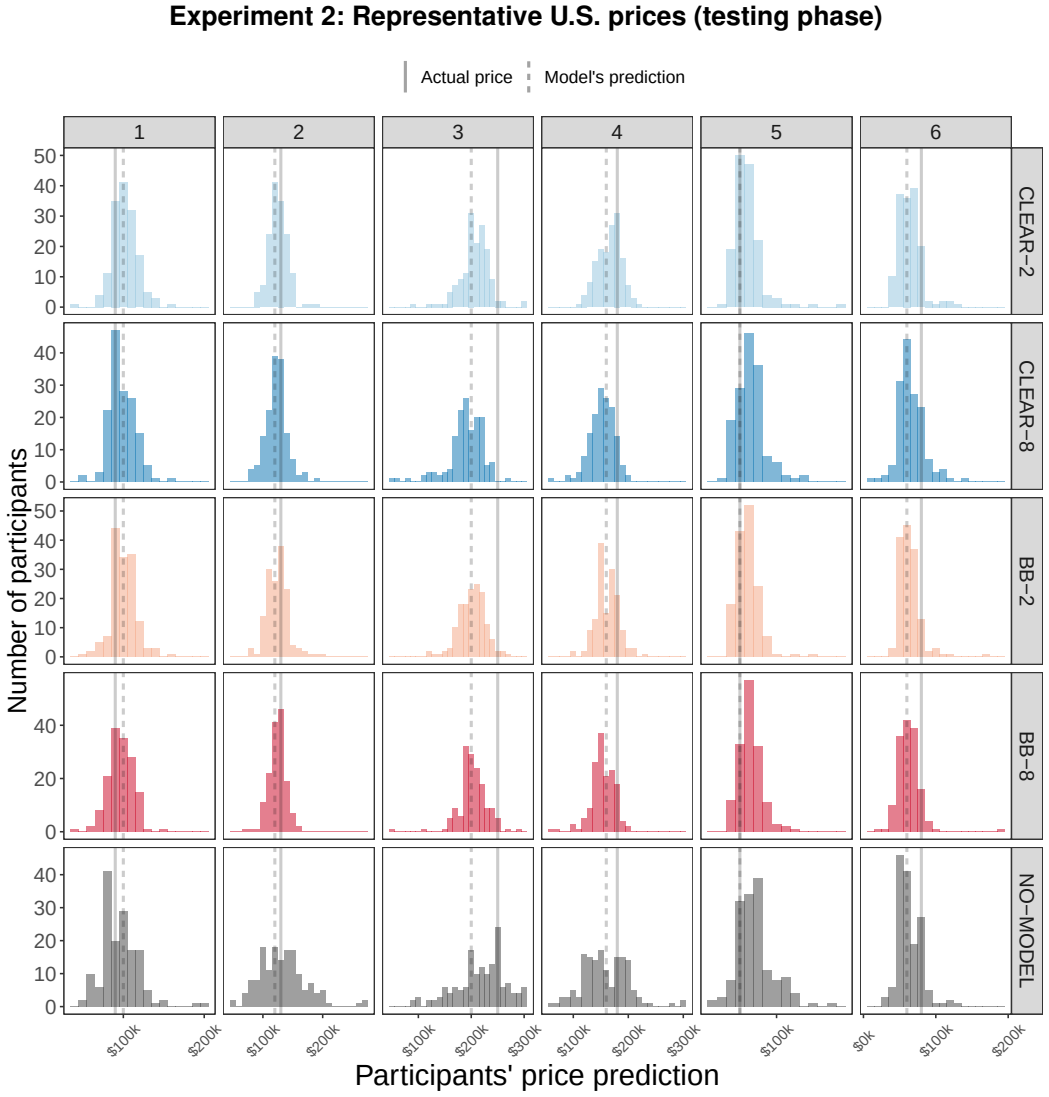


Fig. 17. Distribution of participants' predictions of prices of apartments 1–6 in the testing phase in Experiment 2.

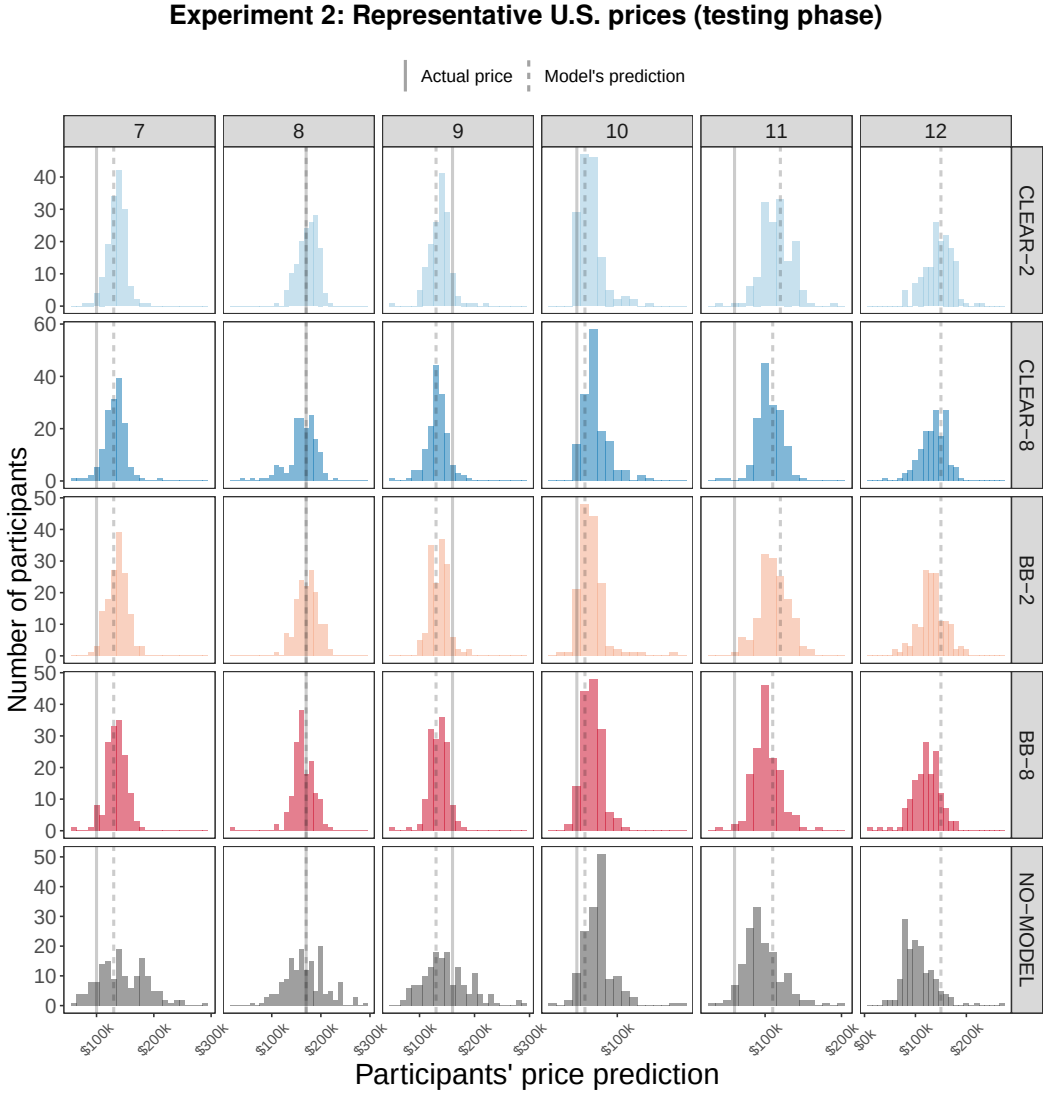


Fig. 18. Distribution of participants' predictions of prices of apartments 7–12 in the testing phase in Experiment 2.

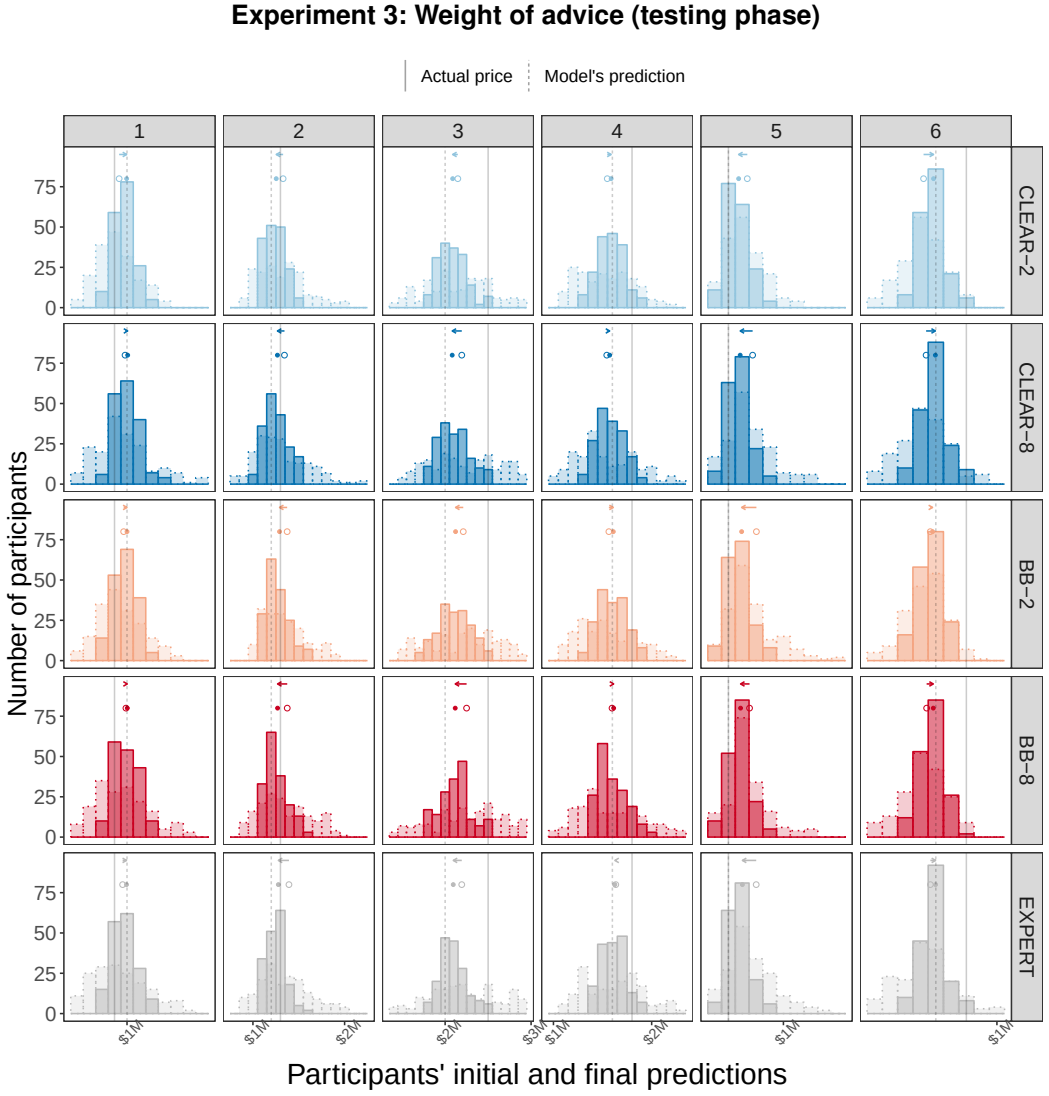


Fig. 19. Distribution of participants' initial (before seeing the model's prediction, dotted) predictions and their final (after seeing the model's prediction, solid) prediction of prices of apartments 1–6 in Experiment 3. Points show the mean initial and final predictions and the arrow indicates the shift in the mean predictions.

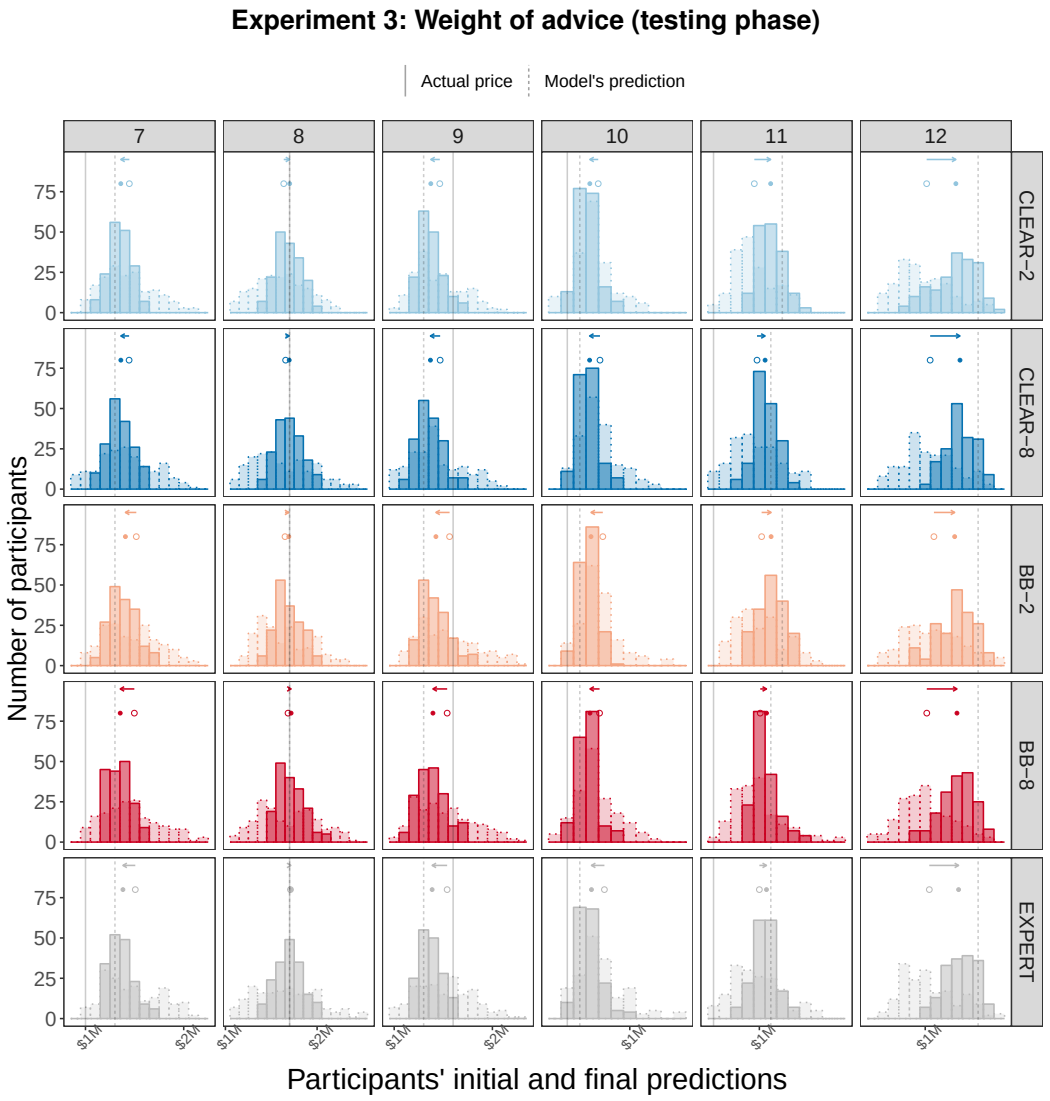


Fig. 20. Distribution of participants' initial (before seeing the model's prediction, dotted) predictions and their final (after seeing the model's prediction, solid) prediction of prices of apartments 7–12 in Experiment 3. Points show the mean initial and final predictions and the arrow indicates the shift in the mean predictions.

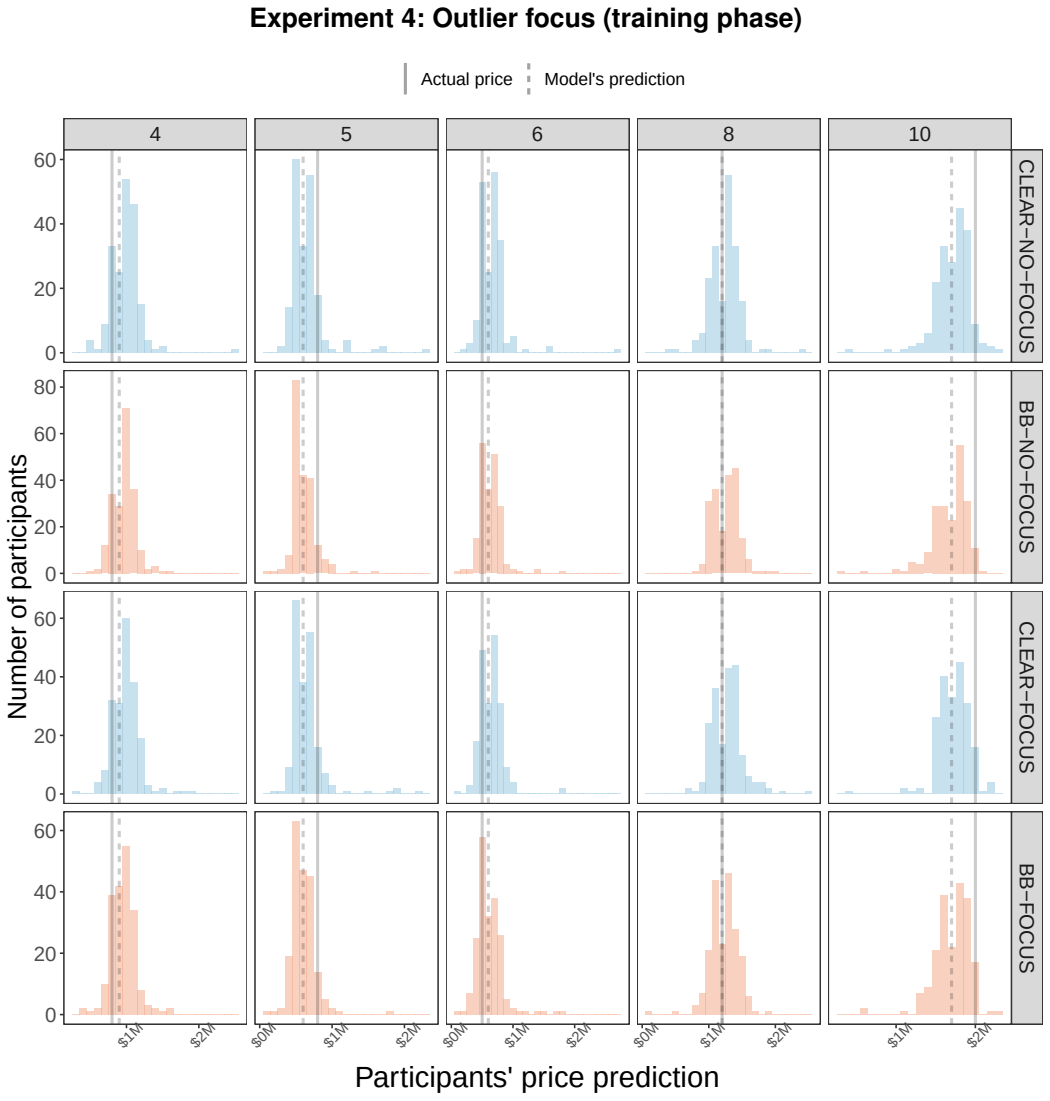


Fig. 21. Distribution of participants' predictions of prices of apartments in the training phase in Experiment 4.

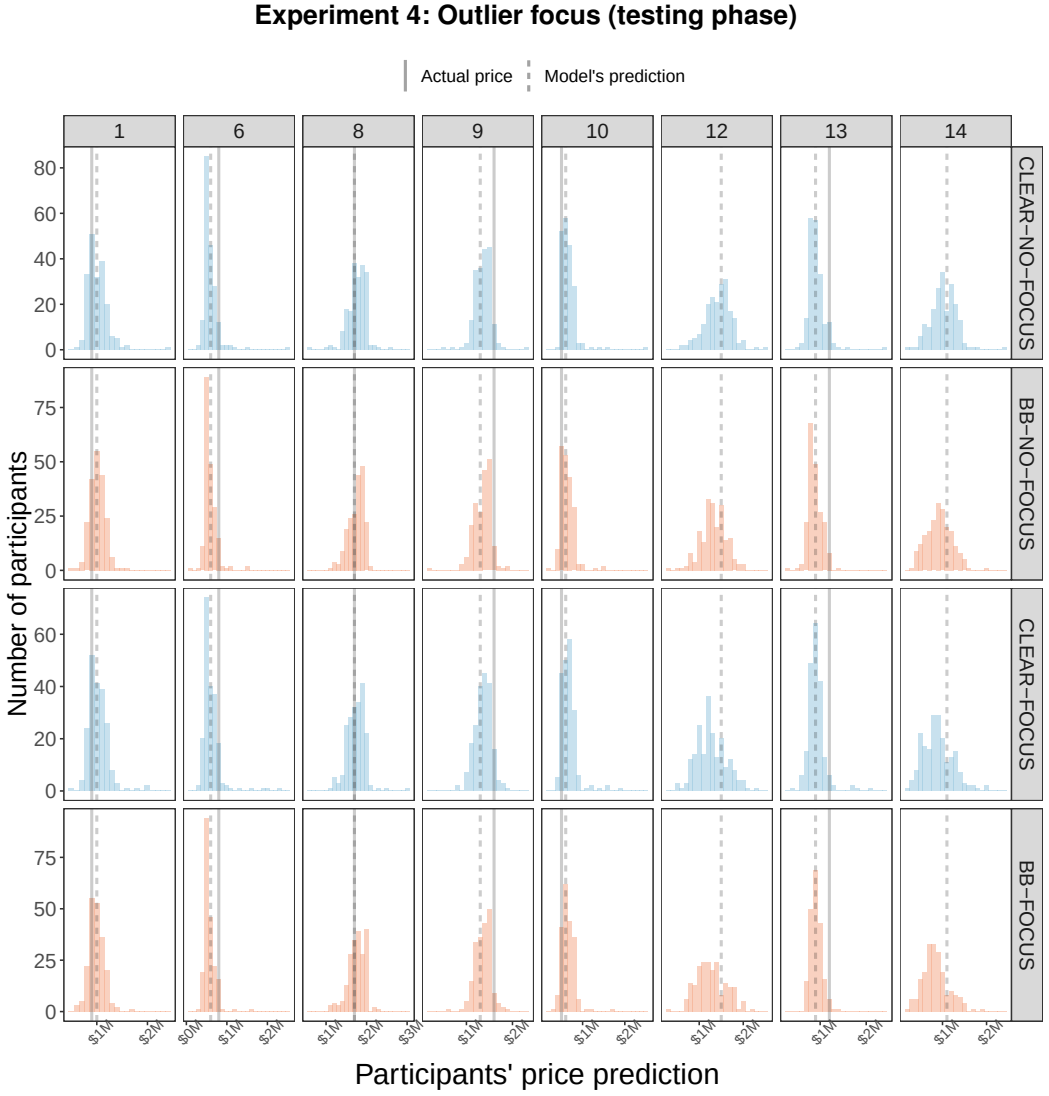


Fig. 22. Distribution of participants' predictions of prices of apartments in the testing phase in Experiment 4.

Appendix F ANOVA TABLES

F.1 Experiment 1: Predicting Prices

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.52	0.52	1.00	994.00	12.57	0.0004
num_features	4.90	4.90	1.00	994.00	119.54	0.0000
transparency:num_features	1.70	1.70	1.00	994.00	41.48	0.0000

Table 6. Results from two-way ANOVA on the simulation error in Experiment 1.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.10	0.10	1.00	994.00	5.83	0.0159
num_features	0.04	0.04	1.00	994.00	2.15	0.1427
transparency:num_features	0.00	0.00	1.00	994.00	0.06	0.8143

Table 7. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price in Experiment 1.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1	0.02	0.02	1.38	0.2405
num_features	1	0.03	0.03	1.70	0.1920
transparency:num_features	1	0.00	0.00	0.00	0.9509
Residuals	994	17.01	0.02		

Table 8. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 11 in Experiment 1.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1	0.34	0.34	8.81	0.0031
num_features	1	0.04	0.04	1.13	0.2882
transparency:num_features	1	0.13	0.13	3.32	0.0687
Residuals	994	38.07	0.04		

Table 9. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 12 in Experiment 1.

F.2 Experiment 2: Scaled-down prices

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.003	0.003	1.00	594.00	7.54	0.0062
num_features	0.032	0.032	1.00	594.00	75.45	0.0000
transparency:num_features	0.018	0.018	1.00	594.00	43.14	0.0000

Table 10. Results from two-way ANOVA on the simulation error in Experiment 2.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.001	0.001	1.00	594.00	3.33	0.0685
num_features	0.000	0.000	1.00	594.00	1.29	0.2556
transparency:num_features	0.000	0.000	1.00	594.00	1.82	0.1775

Table 11. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price in Experiment 2.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1.0	0.001	0.001	3.29	0.0702
num_features	1.0	0.001	0.001	4.51	0.0340
transparency:num_features	1.0	0.000	0.000	1.20	0.2731
Residuals	594.0	0.092	0.000		

Table 12. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 11 in Experiment 2.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1.0	0.007	0.007	17.53	0.0000
num_features	1.0	0.002	0.002	4.05	0.0446
transparency:num_features	1.0	0.001	0.001	2.31	0.1291
Residuals	594.0	0.229	0.000		

Table 13. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 12 in Experiment 2.

F.3 Experiment 3: Weight of Advice

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	0.07	0.07	1.00	798.00	4.20	0.0409
num_features	0.01	0.01	1.00	798.00	0.66	0.4151
transparency:num_features	0.02	0.02	1.00	798.00	1.20	0.2731

Table 14. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price in the four primary conditions in Experiment 3.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
condition	0.09	0.02	4.00	994.00	1.45	0.2147

Table 15. Results from one-way ANOVA on the deviation between the model's prediction and participants' prediction of the price in all conditions (including the "human expert" condition) in Experiment 3.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
transparency	2.14	2.14	1.00	817.77	10.47	0.0013
num_features	0.42	0.42	1.00	817.77	2.07	0.1509
transparency:num_features	0.32	0.32	1.00	817.77	1.57	0.2109

Table 16. Results from two-way ANOVA on the weight of advice in the four primary conditions in Experiment 3.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
condition	2.92	0.73	4.00	1013.65	3.77	0.0048

Table 17. Results from one-way ANOVA on weight of advice in all conditions (including the "human expert" condition) in Experiment 3.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1	0.01	0.01	0.92	0.3380
num_features	1	0.15	0.15	10.87	0.0010
transparency:num_features	1	0.03	0.03	2.08	0.1497
Residuals	798	10.86	0.01		

Table 18. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 11 in Experiment 3.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transparency	1	0.00	0.00	0.04	0.8439
num_features	1	0.38	0.38	10.26	0.0014
transparency:num_features	1	0.01	0.01	0.35	0.5516
Residuals	798	29.32	0.04		

Table 19. Results from two-way ANOVA on the deviation between the model's prediction and participants' prediction of the price for apartment 12 in Experiment 3.