Manipulating and Measuring Model Interpretability

Forough Poursabzi-Sangdeh
forough.poursabzi@microsoft.com
Microsoft Research

Daniel G. Goldstein
dgg@microsoft.com
Microsoft Research

Jake M. Hofman
jmh@microsoft.com
Microsoft Research

Jennifer Wortman Vaughan
jenn@microsoft.com
Microsoft Research

Hanna Wallach
wallach@microsoft.com
Microsoft Research

Abstract

Despite a growing literature on creating interpretable machine learning methods, there have been few experimental studies of their effects on end users. We present a series of large-scale, randomized, pre-registered experiments in which participants were shown functionally identical models that varied only in two factors thought to influence interpretability: the number of input features and the model transparency (clear or black-box). Participants who were shown a clear model with a small number of features were better able to simulate the model's predictions. However, contrary to what one might expect when manipulating interpretability, we found no significant difference in multiple measures of trust across conditions. Even more surprisingly, increased transparency hampered people's ability to detect when a model has made a sizeable mistake. These findings emphasize the importance of studying how models are presented to people and empirically verifying that interpretable models achieve their intended effects on end users.

1 Introduction

Machine learning is increasingly used to make decisions that affect people's lives in critical domains like criminal justice, fair lending, and medicine. Machine learning models are often evaluated based on their predictive performance on held-out data sets, measured, for example, in terms of accuracy, precision, or recall. However, good performance on held-out data may not be sufficient to convince decision makers that a model is trustworthy or reliable in the wild.

To address this problem, a new line of research has emerged that focuses on developing *interpretable* machine learning methods. There are two common approaches. The first is to employ *simple* models in which the impact of each feature on the model's prediction is arguably easy to understand. Examples include generative additive models [3, 18, 19] and point systems [10, 27]. The second is to provide post-hoc *explanations* for (potentially complex) models. One thread of research in this direction looks at how to explain individual predictions by learning simple local approximations of a model around particular data points [14, 20, 24] or estimating the influence of training examples [11], while another focuses on visualization of model output or properties [12, 28].

Despite the progress in this area, there is still no consensus about how to define, quantify, or measure the interpretability of a machine learning model [5]. Different notions of interpretability, such as simulatability, trustworthiness, and simplicity, are often conflated [16]. This problem is exacerbated by the fact that there are different types of users of machine learning systems and these users may have different needs in different scenarios [25]. The approach that works best for a regulator who wants to understand why a particular person was denied a loan may be different from the approach that works best for a data scientist debugging a machine learning model or a CEO using a model to make a high-stakes decision.

We take the perspective that the difficulty of defining interpretability stems from the fact that interpretability is not something that can be directly manipulated or measured. Rather, interpretability is a latent 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 outcomes* (such as an end user's ability to simulate, trust, or debug the model). Different factors may influence these outcomes in different ways. As such, we argue that to understand interpretability, it is necessary to directly manipulate and measure the influence that different factors have on people's abilities to complete various tasks.

This endeavor goes beyond the realm of typical machine learning research, extending into human-computer interaction. While the factors that influence interpretability are properties of the system design, the outcomes that we would ultimately like to measure are properties of human behavior. Because of this, building interpretable machine learning models is not a purely computational problem. In other words, what is or is not "interpretable" is defined by people, not algorithms.

We therefore take a human-centered approach, building on previous psychology and social science research on human trust in models [4, 17, 21] and similar efforts in the machine learning and human-computer interaction communities [2, 6, 8, 13, 15, 23, 24]. We present a sequence of large-scale, randomized human-subject experiments in which we vary factors thought to make models more or less interpretable and measure how these changes impact people's decision making. Based on an extensive review of the literature on interpretable machine learning, we focus on two factors that are often assumed to influence interpretability, but rarely studied formally: the number of features and the model transparency, i.e., whether the model internals are *clear* or a *black box*. Similarly, we focus on three outcomes commonly mentioned in this literature: simulatability, trust (measured in terms of deviation from the model's predictions), and the ability to detect when the model makes a mistake. We focus on housing price prediction as a domain since it should have familiar features (e.g., numbers of bedrooms and bathrooms) and be interesting to participants, as many people have purchased or considered purchasing a home.

We investigate the effect of factors associated with interpretability through three main questions:

- **How well can people estimate what a model will predict?** Here we find that people can better simulate a model's predictions when presented with a clear model with few features compared to other experimental conditions.
- **How much do people trust (follow) a model's predictions?** Contrary to what one might expect when manipulating factors associated with interpretability, we do not find a significant difference in trust across conditions.
- **How well can people detect when a model has made a sizeable mistake?** Even more surprisingly, we find that transparency can hamper people's ability to detect when a model makes serious mistakes.

In each of our pre-registered experiments, 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. 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 model prediction for each apartment, which came from a linear regression model. What varied between the experimental conditions was *only the presentation of the model*. As a result, any observed differences in the participants' behavior between conditions could be attributed entirely to the model presentation. This is a key feature of our experimental design.

In our first and primary experiment, we showed people a sequence of 12 apartments. The first 10 apartments had typical configurations whereas the last two had unusual combinations of features (e.g., more bathrooms than bedrooms). For each apartment, participants first saw its features alongside the model and were asked to estimate what the model would predict for the apartment's selling price. They were then shown the model's prediction and asked for their own estimate of what the apartment sold for.

We hypothesized that participants who were shown a clear model with a small number of features would be better able to simulate the model's predictions and more likely to trust (and thus follow) the model's predictions. We also hypothesized that participants in different conditions would exhibit varying abilities to correct the model's inaccurate predictions on unusual apartments. As predicted, participants who were shown a clear model with a small number of features were better able to simulate the model's predictions; however, we did not find that they were more likely to follow the model's predictions and instead found no difference in trust between conditions for the typical apartments. We also found that participants who were shown a clear model were less able to correct the model's inaccurate predictions on the unusual apartments.

We ran three additional experiments to better understand these results and check their robustness to our experimental design.

In our second experiment, we scaled down the apartment prices and maintenance fees to match median housing 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. Even with scaled-down prices and fees, the findings from our first experiment replicated.

We designed our third experiment to check whether these results would change under a different measure of trust. Instead of quantifying trust by how close people's predictions were to the model's predictions, we used an alternative metric known as *weight of advice*, commonly used in the literature on advice-taking [7, 29] and subsequently used in the context of algorithmic predictions by Logg [17]. We calculated weight of advice by eliciting two predictions from participants: one before being introduced to the model and one after. We again found no difference in trust between conditions. Interestingly, however, participants in the clear conditions were better able to correct the model's mistakes than in the previous two experiments. One possible reason for this improvement is that making predictions before being exposed to the clear model's overwhelming level of detail increased the likelihood that participants noticed unusual combinations of features and lowered their estimates accordingly.

This motivated our fourth and final experiment, which focused on whether the observed differences in error detection between conditions could be explained by people failing to notice the unusual apartments in the clear conditions. We returned to the setup for the first experiment but added a condition in which some people were shown an explicit "attention check" to make them aware of peculiar configurations. Without the attention check, we again found that participants who were shown a clear model were less able to correct inaccurate predictions on examples for which the model made a sizeable mistake. However, this difference essentially disappeared when participants were provided with the attention check, consistent with the idea that transparency can be overwhelming and cause users to overlook unusual cases.

Across several experiments we see that factors thought to improve interpretability can have negligible effects on typical cases and even detrimental effects on unusual ones. While a researcher designing a machine learning model might expect that exposing model internals would increase people's ability to detect when the model will make a mistake, we find the opposite to be true across multiple studies. Taken together, these results emphasize the importance of user testing over intuition in the design of interpretable machine learning methods.

In the remainder of the paper we provide further details for each of these experiments and present our results in greater depth. We conclude by discussing limitations of and possible extensions to our work, as well as implications for the design of interfaces to machine learning models.

2 Experiment 1: Predicting apartment prices

Our first experiment was designed to measure the influence of the number of features and model transparency (clear or black box) on three tasks that our literature search revealed to be commonly associated with interpretability: laypeople's abilities to simulate a model's predictions, gain trust in a model, and understand when a model makes mistakes. Before running the experiment, we posited and pre-registered three hypotheses, stated informally here:[1]

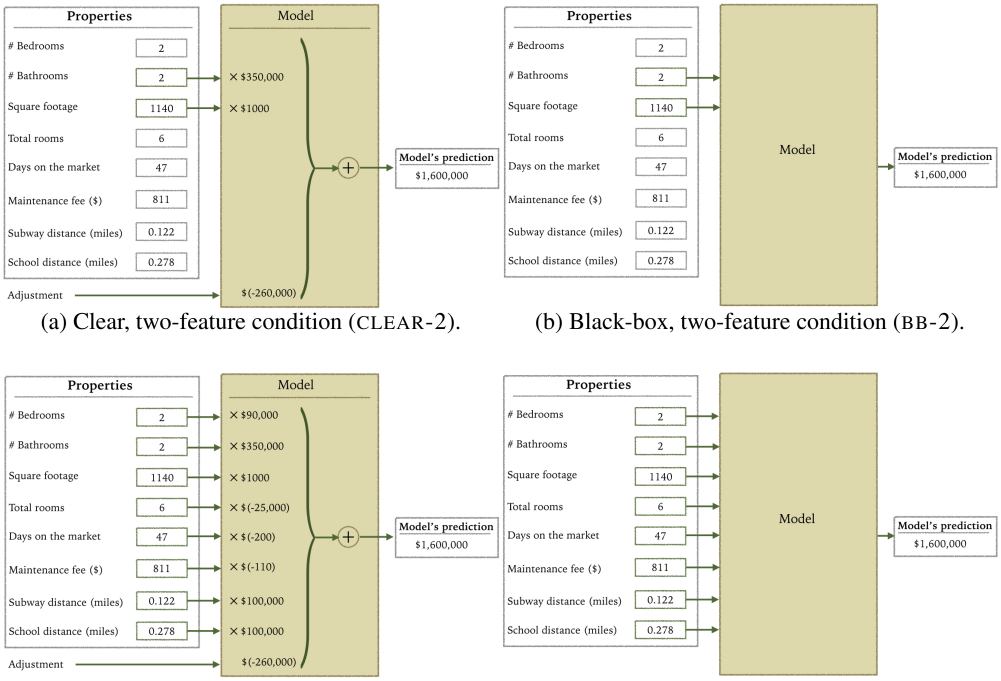
H1. **Simulation.** A clear model with a small number of features will be easiest for participants to simulate.
H2. **Trust.** Participants will be more likely to trust (and thus follow) the predictions of a clear model with a small number of features than the predictions of a black-box model with a large number of features.
H3. **Detection of mistakes.** Participants in different conditions will exhibit varying abilities to correct the model's inaccurate predictions on unusual examples.

We test the first hypothesis by showing people an apartment configuration, asking them to estimate what the model will predict for its selling price, and comparing this estimate with the model's prediction. A small difference between these two quantities indicates that people have a good understanding of how the model works. For the second hypothesis, we quantify trust by showing people the model's prediction for each apartment, asking them to estimate the actual selling price, and measuring the deviation between this estimate and the model's prediction. Small deviations from the model's predictions indicate high trust in the model, whereas large deviations imply that people don't trust the model. We use the same deviation measure for the third hypothesis, but applied to unusual apartments for which the model makes erroneously high predictions. Here a large deviation from the model's predictions implies that people are able to correct the model's mistakes.

For the unusual examples, we intentionally did not pre-register any hypotheses about which conditions would make participants more or less able to correct inaccurate predictions. On the one hand, if a participant understands the model better, she may be better equipped to correct examples on which the model makes mistakes. On the other hand, a participant may place greater trust in a model she understands well, leading her to closely follow its predictions.

We additionally pre-registered our intent to analyze participants' prediction error in each condition, but intentionally did not pre-register any directional hypotheses.

[1] Pre-registered hypotheses are available at `https://aspredicted.org/xy5s6.pdf`.



(a) Clear, two-feature condition (CLEAR-2).

(b) Black-box, two-feature condition (BB-2).

(c) Clear, eight-feature condition (CLEAR-8).

(d) Black-box, eight-feature condition (BB-8).

Figure 1: The four primary experimental conditions. In the conditions on top, the model used two features; on the bottom, it used eight. In the conditions on the left, participants saw the model internals; on the right, they were presented with the model as a black box.

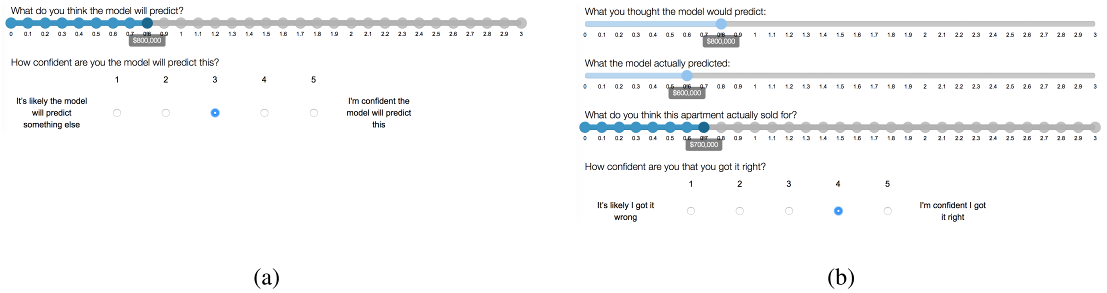
2.1 Experimental design

As explained in the previous section, we asked participants to predict apartment prices with the help of a machine learning model. We showed all participants the same set of apartments and the same model prediction for each apartment, regardless of their randomly assigned experimental condition. This key feature of our experimental design is what enabled us to run tightly controlled experiments. *The only thing that varied between the primary conditions was model presentation*. We considered four primary conditions in a 2×2 design:

- Participants were randomly assigned to see either a model that uses only two features (number of bathrooms and square footage—the two most predictive features) or a model that uses all eight features. (Note that all eight feature values were visible to participants in all conditions.)
- Participants were randomly assigned to either see the model internals (i.e., a linear regression model with visible coefficients) or see the model as a black box.

Screenshots from each of the four primary experimental conditions are shown in Figure 1. We additionally considered a baseline condition in which there was no model available.

We chose to use the two most predictive features in the two-feature model because this resulted in the most reasonable



What do you think the model will predict?

$800,000

How confident are you the model will predict this?

It's likely the model will predict something else	1	2	3	4	5	I'm confident the model will predict this
	○	○	●	○	○	

(a)

What you thought the model would predict:

$800,000

What the model actually predicted:

$600,000

What do you think this apartment actually sold for?

$700,000

How confident are you that you got it right?

It's likely I got it wrong	1	2	3	4	5	I'm confident I got it right
	○	○	○	●	○	

(b)

Figure 2: Part of the testing phase in the first experiment: (a) participants were asked to guess the model's prediction and state their confidence (step 1) and (b) participants were asked to make their own prediction and state their confidence (step 3).

and realistic two-feature model possible. The accuracies of the two- and eight-feature models were very close on training data, as described below, and their (rounded) predictions were identical on all apartments used in the experiment.

We ran the experiment on Amazon Mechanical Turk using psiTurk [9], an open-source platform for designing online experiments. The experiment was IRB-approved. We recruited 1,250 participants, all located in the U.S., with Mechanical Turk approval ratings greater than 97%. The participants were randomly assigned to the five conditions (CLEAR-2, $n = 248$; CLEAR-8, $n = 247$; BB-2, $n = 247$; BB-8, $n = 256$; and NO-MODEL, $n = 252$) and each participant received a flat payment of \$2.50.

Participants were first shown detailed instructions (including, in the clear conditions, a simple English description of the corresponding two- or eight-feature linear regression model), before proceeding with the experiment in two phases. The *training phase* familiarized participants with both the housing domain 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 price, asked to make their own prediction, and then shown the apartment's actual price. In the baseline condition, participants were asked to predict the price of each apartment and then shown the actual price.

In the *testing phase*, participants were shown another twelve apartments. The order of the first ten was randomized, while the remaining two always appeared last, for reasons described below. In the four primary experimental conditions, participants were asked to guess what the model would predict for each apartment (i.e., simulate the model) and to indicate how confident they were in this guess on a five-point scale (Figure 2a). They were then shown the model's prediction and asked to indicate how confident they were that the model was correct. Finally, they were asked to make their own prediction of the apartment's price and to indicate how confident they were in this prediction (Figure 2b). In the baseline condition, participants were asked to predict the price of each apartment and to indicate their confidence.

The apartments shown to participants were selected from a data set of actual Upper West Side apartments taken from StreetEasy.com, a popular and reliable New York City real estate website, between 2013 and 2015. To create the models for the four primary experimental conditions, we first trained a two-feature linear regression model on our data set using ordinary least squares with Python's scikit-learn library [22], rounding coefficients to "nice" numbers within a safe range.[2] To keep the models as similar as possible, we fixed the coefficients for number of bathrooms and square

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

footage and the intercept of the eight-feature model to match those of the two-feature model, and then trained a linear regression model with the remaining six features, following the same rounding procedure to obtain “nice” numbers. The resulting coefficients are shown in Figure 1. When presenting the model predictions to participants, we rounded predictions to the nearest $100,000.

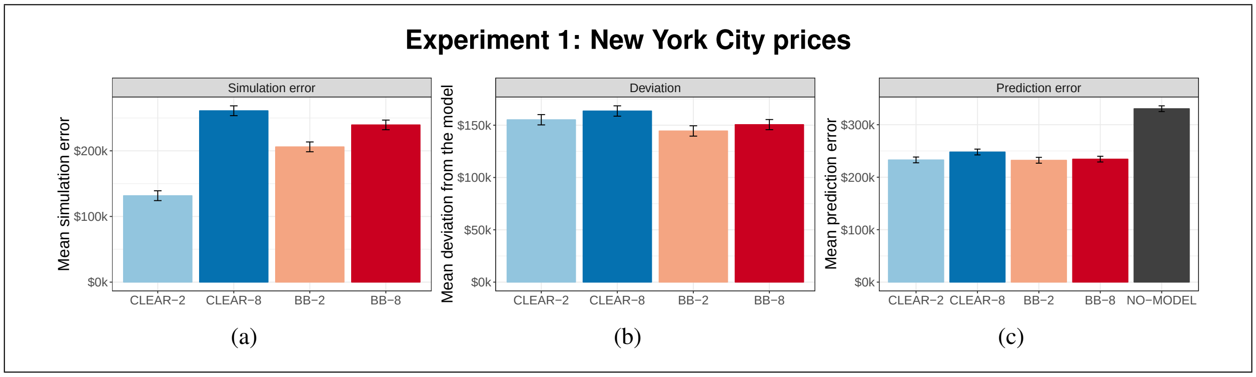
We used the same set of apartments for each participant since randomizing the selection of apartments would introduce additional noise and reduce the power of the experiments, making it harder to spot differences between conditions. To enable comparisons across experimental conditions, the ten apartments used in the training phase and the first ten apartments used in the testing phase were selected from those apartments in our data set for which the rounded predictions of the two- and eight-feature models agreed and chosen to cover a wide range of deviations between the models' predictions and the apartments' actual prices. By selecting only apartments for which the two- and eight-feature models agreed, we were able to ensure that what varied between conditions was only the presentation of the model. As a result, any observed differences in the participants' behavior between conditions could be attributed entirely to the model presentation.

The last two apartments used in the testing phase were chosen to test our third hypothesis—i.e., that participants in different conditions will exhibit varying abilities to correct the model's inaccurate predictions on unusual examples. To test this hypothesis, we would ideally have used an apartment with strange or misleading features that caused the two- and eight-feature models to make the same bad prediction. Unfortunately, there was no such apartment in our data set, so we chose two examples to test different aspects of our hypothesis. Both of these examples exploited the models' large coefficient ($350,000) for number of bathrooms. The first (apartment 11) was a one-bedroom, two-bathroom apartment from our data set for which both models made high, but different, predictions. Comparisons between the two- and eight- feature conditions were therefore impossible, but we could examine differences in accuracy between the clear and black-box conditions. The second (apartment 12) was a synthetically generated one-bedroom, three-bathroom apartment for which both models made the same (high) prediction, allowing comparisons between all conditions, but ruling out accuracy comparisons since there was no ground truth. These apartments were always shown last to avoid the previously studied phenomenon in which people trust a model less after seeing it make a mistake [4].

2.2 Results

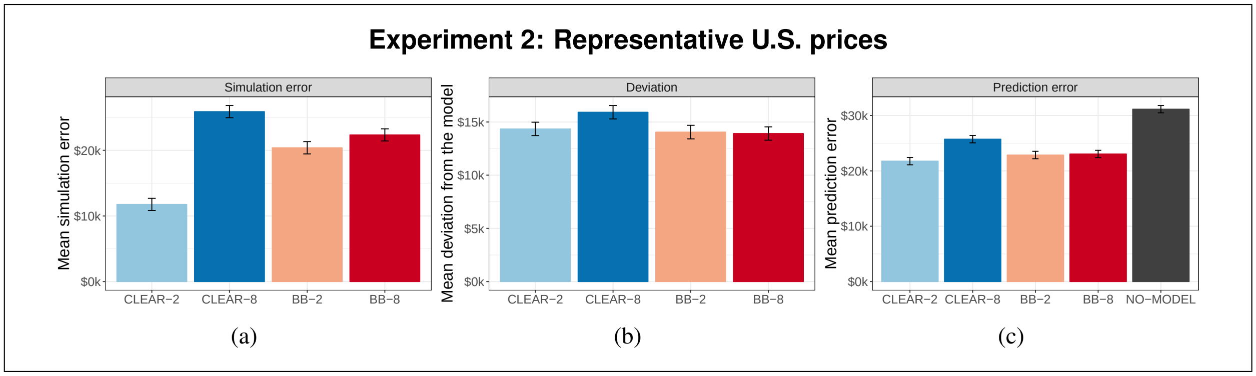
Having run our experiment, we compared participants' behavior across the conditions. Doing so required us to compare multiple responses from multiple participants, which was complicated by possible correlations among any given participant's responses. For example, some people might consistently overestimate apartment prices regardless of the condition they are assigned to, while others might consistently provide underestimates. We addressed this by fitting a mixed-effects model for each measure of interest to capture differences across conditions while controlling for participant-level effects—a standard approach for analyzing repeated measures experimental designs [1]. We derived all plots and statistical tests from these models; plots show averages with one standard error by condition from the fitted models, and statistical tests report degrees of freedom, test statistics, and p-values under the models. Unless otherwise noted, all plots and statistical tests correspond to just the first ten apartments from the testing phase.

H1. **Simulation.** We defined a participant's simulation error to be the absolute deviation between the model's prediction, m, and the participant's guess for that prediction, u_m—that is, $|m - u_m|$. Figure 3a shows the mean simulation error in the testing phase. As hypothesized, participants in the CLEAR-2 condition had lower simulation error, on average, than participants in the other conditions ($t(996) = 11.91, p < .001$). This suggests that, on average, participants in this



(a) (b) (c)

Figure 3: Results from our first experiment: (a) mean simulation error, (b) mean deviation of participants' predictions from the model's prediction (a smaller value indicates higher trust), and (c) mean prediction error. Error bars indicate one standard error.



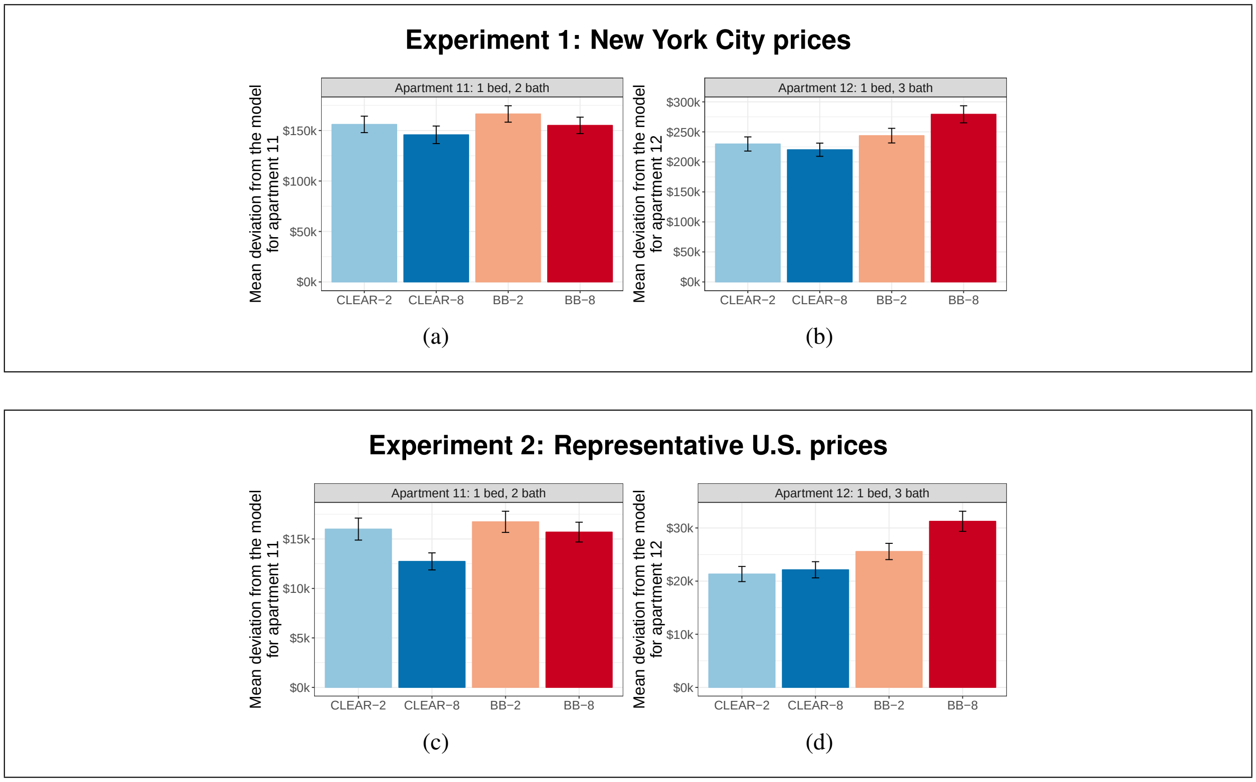
(a) (b) (c)

Figure 4: Results from our second experiment, which replicate the findings of our first experiment.

condition had better understanding of how the model works. We also found that participants were better able to simulate the predictions of a model that uses two features than a model that uses eight features ($t(996) = 10.65$, $p < .001$ for the contrast of CLEAR-2 and BB-2 with CLEAR-8 and BB-8). Although we had not pre-registered this as a hypothesis, we find it interesting. Participants in the CLEAR-8 condition appeared to have *higher* simulation error, on average, than participants in the BB-8 condition who could not see the model's internals ($t(996) = 3.00$, $p = .002$), though we note that this comparison is not one we pre-registered and could be due to chance.

H2. **Trust.** To measure trust, we calculated the absolute deviation between the model's prediction, m, and the participant's prediction of the apartment's price, u_a—that is, $|m - u_a|$; a smaller value indicates higher trust. Figure 3b shows that contrary to our second hypothesis, we found no significant difference in participants' deviation from the model between CLEAR-2 and BB-8.

H3. **Detection of mistakes.** We used the last two apartments in the testing phase (apartments 11 and 12) to test our



(a)

(b)

(c)

(d)

Figure 5: Mean deviation from the model for apartments 11 and 12 in our first experiment (top) and in our second experiment (bottom). Error bars indicate one standard error. (Note that for apartment 11, comparisons between the two- and eight-feature conditions are not possible because the models make different predictions.)

third hypothesis. The models made erroneously high predictions on these examples. For both apartments, participants in the four primary experimental conditions overestimated the apartments' prices, compared to participants in the baseline condition. We suspect that this is due to an anchoring effect around the models' predictions. For apartment 11, we found no significant difference in participants' deviation from the model's prediction between the four primary conditions (see Figure 5a). For apartment 12, we found a significant difference between the clear and black-box conditions ($t(996) = 2.96$, $p = .003$ for the contrast of CLEAR-2 and CLEAR-8 with BB-2 and BB-8). In particular, participants in the clear conditions deviated from the model's prediction less, on average, than participants in the black-box conditions, resulting in even worse final predictions of the apartment's price (see Figure 5b). One potential concern about this finding is the possibility of cognitive fatigue when participants make predictions for the last two apartments. To test whether this was the case, we fit a linear regression model to participants' simulation error over time. We found that participants' simulation error significantly decreased over time, which indicates that there is no evidence of cognitive fatigue. Participants' poor abilities to detect the model's mistakes in clear conditions contradicts the common intuition that transparency enables users to understand when a model will make mistakes. We explore this finding in more detail later.

Prediction error. We defined prediction error to be the absolute deviation between the apartment's actual price, a, and the participant's prediction of the apartment's price, u_a—that is, $|a - u_a|$. Figure 3c shows that we did not find a significant difference between the four primary experimental conditions. Participants in the primary conditions had higher average error than the model itself, but fared better than those in the baseline condition ($t(1248) = 15.27$, $p < .001$ for the contrast of the baseline with the four primary conditions), indicating that the use of a model is advantageous in this setting.

3 Experiment 2: Scaled-down prices

One potential critique of our first experiment is that participants' lack of familiarity with New York City's unusually high apartment prices may influence their trust in the model and ability to detect when the model makes a mistake. Our second experiment was designed as a robustness check to address this potential concern by replicating our first experiment with apartment prices and maintenance fees scaled down to match median housing prices in the U.S. Before running this experiment we pre-registered three hypotheses.[3] The first two hypotheses (H4 and H5) are identical to H1 and H2 from our first experiment. We made the third hypothesis more precise than H3 to reflect the results of our first experiment and the results of a small pilot with scaled-down prices:

H6. **Detection of mistakes.** Participants will be less likely to correct inaccurate predictions on unusual examples of a clear model compared to a black-box model, and this effect will be more prominent the more unusual an example is.

3.1 Experimental design

We first scaled down the apartment prices and maintenance fees from our first experiment by a factor of ten. To account for this change, we also scaled down all regression 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. We again ran the experiment on Amazon Mechanical Turk. We excluded people who had participated in our first experiment, and recruited 750 new participants all of whom satisfied the selection criteria from our first experiment. The participants were randomly assigned to the five conditions (CLEAR-2, $n = 150$; CLEAR-8, $n = 150$; BB-2, $n = 147$; BB-8, $n = 151$; and NO-MODEL, $n = 152$) and each participant received a flat payment of \$2.50.

3.2 Results

H4. **Simulation.** As hypothesized, and shown in Figure 4a, participants in the CLEAR-2 condition had significantly lower simulation error, on average, than participants in the other primary conditions ($t(596) = 10.28$, $p < .001$). We also found that participants were better able to simulate the predictions of the model that uses two features than the model that uses eight features ($t(596) = 8.39$, $p < .001$ for the contrast of CLEAR-2 and BB-2 with CLEAR-8 and BB-8). Although we had not pre-registered this as a hypothesis, we find it interesting. This is in line with the finding from our first experiment.

[3] Pre-registered hypotheses are available at `https://aspredicted.org/3bv8i.pdf`.

H5. **Trust.** Contrary to our second hypothesis, and in line with the finding from our first experiment, we found no significant difference in participants' trust, as indicated by their deviation from the model, between CLEAR-2 and BB-8.

H6. **Detection of mistakes.** As hypothesized, and in line with our findings from experiment 1, participants in the clear conditions deviated from the model's prediction less, on average, than participants in the black-box conditions for apartment 12, resulting in even worse final predictions of the apartment's price (see Figure 5d, $t(596) = 4.17$, $p < .001$). However, as in experiment 1, we found no significant difference in participants' deviation from the model's prediction between the four primary conditions for apartment 11 (Figure 5c), perhaps because the configuration is not sufficiently unusual.

Prediction error. Participants in the CLEAR-2 condition had statistically but not practically (< \$3,000) significantly lower prediction error than participants in the other primary conditions ($t(596) = 3.17$, $p = .001$).

In summary, results from our first experiment replicated with the scaled-down prices, which suggests that the lack of familiarity with New York City's high housing prices does not explain participants' poor ability in detecting the model's mistakes in clear conditions.

4 Experiment 3: Alternative measure of trust

In our first two experiments, participants were no more likely to trust 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, as indicated by the deviation of their own predictions from the model's prediction. However, perhaps another measure of trust would reveal differences between the conditions. In this section, we therefore present our third experiment, which was designed to allow us to compare participants' trust across the conditions using an alternative measure of trust: the *weight of advice* measure frequently used in the literature on advice-taking [7, 17, 29].

Weight of advice quantifies the degree to which people update their beliefs (e.g., predictions made *before* seeing the model's predictions) toward advice they are given (e.g., the model's predictions). In the context of our experiment, it is defined as $|u_2 - u_1| \,/\, |m - u_1|$, where m is the model's prediction, u_1 is the participant's initial prediction of the apartment's price before seeing m, and u_2 is the participant's final prediction of the apartment's price after seeing m. It 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 comparing weight of advice across the conditions, consider the scenario in which u_2 is close to m. There are different reasons why this might happen. On the one hand, it could be the case that u_1 was far from m and the participant made a significant update to their initial prediction based on the model. On the other hand, it could be the case that u_1 was already close to m and the participant did not update her prediction at all. These two scenarios are indistinguishable in terms of the participant's deviation from the model's prediction. In contrast, weight of advice would be high in the first case and low in the second.

We additionally used this experiment to check whether participants' behavior would differ 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 [4, 21]. Most closely related to our experiment, Logg [17] found that when people were presented with predictions from either an algorithm or a human expert, they updated their own predictions toward those of an algorithm more than they did toward those of a human expert in a variety of domains. We were interested to see whether this finding would replicate. We pre-registered four hypotheses:[4]

H7. **Trust (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. **Trust (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 will trust a human expert and a black-box model with a large number of features to differing extents. As a result, their deviation from the model's predictions and their weight of advice will also differ.

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.

4.1 Experimental design

We returned to using the original New York City housing prices and considered the same four primary experimental conditions as in the first two experiments plus a new condition, EXPERT, in which participants saw the same information as in BB-8, but with the black-box 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 participants to predict apartment prices (i.e., the first step of the testing phase described below).

We again ran the experiment on Amazon Mechanical Turk. We excluded people who had participated in our first two experiments, and recruited 1,000 new participants all of whom satisfied the selection criteria from our first two experiments. The participants were randomly assigned to the five conditions (CLEAR-2, $n = 202$; CLEAR-8, $n = 200$; BB-2, $n = 202$; BB-8, $n = 198$; and EXPERT, $n = 197$) and each participant received a flat payment of \$1.50. We excluded data from one participant who reported technical difficulties.

We asked participants to predict apartment prices for the same set of apartments used in the first two experiments. However, in order to calculate weight of advice, we modified the experiment design so that participants were asked for two predictions for each apartment during the testing phase: an initial prediction before being shown the model's prediction and a final prediction after being shown the model's prediction. To ensure that participants' initial predictions were the same across the conditions, we asked for their initial predictions for all twelve apartments before introducing them to the model or human expert and before informing them that they would be able to update their predictions. This design has the added benefit of potentially reducing the amount of anchoring on the model or expert's predictions.

Participants were first shown detailed instructions (which intentionally did not include any information about the corresponding model or human expert), before proceeding with the experiment in two phases. In the (short) training phase, participants were shown three apartments, asked to predict each apartment's price, and shown the apartment's actual price. The testing phase consisted of two steps. In the first step, participants were shown another twelve apartments. The order of all twelve apartments was randomized. Participants were asked to predict the price of each apartment. In the second step, participants were introduced to the model or human expert before revisiting the twelve apartments. As

[4] Pre-registered hypotheses are available at `https://aspredicted.org/795du.pdf`.

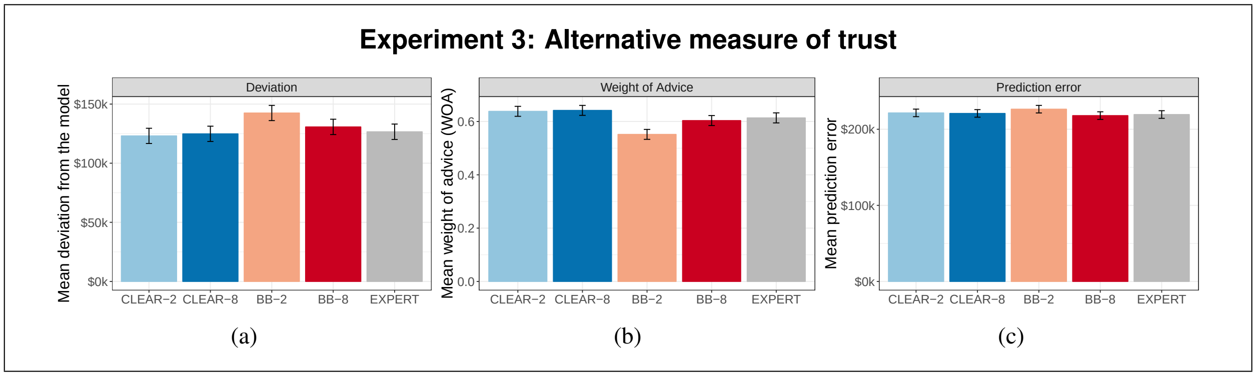


Figure 6: Results from our third experiment: (a) mean deviation of participants' predictions from the model's prediction (a smaller value indicates higher trust), (b) mean weight of advice, and (c) mean prediction error. Error bars indicate one standard error.

in the first two experiments, the order of the first ten apartments was randomized, while the remaining two (apartments 11 and 12) always appeared last. For each apartment, participants were first reminded of their initial prediction, then shown the model or expert's prediction, and then asked to make their final prediction of the apartment's price.[5]

4.2 Results

H7. **Trust (deviation).** Contrary to our first hypothesis, and in line with the findings from our first two experiments, we found no significant difference in participants' deviation from the model between CLEAR-2 and BB-8 (see Figure 6a). H8. **Trust (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 weight of advice over all participant–apartment pairs for which the participant's initial prediction did not match the model's prediction.[6] This calculation can be viewed as calculating the mean conditioned on there being initial disagreement between the participant and the model. Contrary to our second hypothesis, and in line with the findings for the measures of trust in our first two experiments, we did not find a significant difference in participants' weight of advice between the CLEAR-2 and BB-8 conditions (see Figure 6b).

H9. **Humans vs. machines.** Contrary to our third hypothesis, we did not find a significant difference in participants' trust, as indicated by either the deviation of their predictions from the model or expert's prediction or by their weight of advice, between the BB-8 and EXPERT conditions.

[5] We initially considered an alternative design in which participants were asked to predict each apartment's price, shown the model's prediction, and then asked to update their own prediction before moving on to the next apartment. During pilots, it appeared that participants changed their initial predictions in response to the model. To verify this, we ran a larger version of this experiment, hypothesizing that participants' initial predictions would deviate less from the model's predictions in the CLEAR-2 condition (`https://aspredicted.org/zi8yy.pdf`). As predicted, this was indeed the case ($t(241) = -3.41$, $p < .001$). The amount by which participants' initial predictions change based on the model they see could be viewed as another measure of trust.

[6] We found no significant difference in the fraction of times that participants' initial predictions matched the model's predictions.

H10. **Detection of mistakes.** In contrast to our first two experiments, we did not find that participants in the clear conditions were less able to correct inaccurate predictions.

5 Experiment 4: Attention check

Recall that participants in the clear conditions in our first two experiments were less likely to notice when the model overpriced an apartment compared with black-box conditions (Figures 5b and 5d). In seeming contradiction, our third experiment revealed no difference across conditions in terms of participants' abilities to detect the model's sizeable mistakes.

We propose an explanation for these findings and support it with an additional experiment. The explanation rests on two conjectures. First, visual displays with a great amount of numerical information (e.g., model coefficients) might cause users not to notice peculiarities of a case under consideration. Second, when past predictions are visible, they might exert an anchoring effect [26] on subsequent predictions, swaying new predictions in the direction of the past predictions.

In the first two experiments, participants made their final prediction while seeing their *simulation of the model's prediction* (Figure 2b). In contrast, participants in the third experiment made their final prediction while seeing their *own initial guess of the price*. That is, there were different anchor values in the first and second experiments compared with the third.

Furthermore, within the first two experiments, anchor values differed by condition because they were influenced by the model presentation. Participants in the CLEAR-2 condition could simulate the model rather well (Figures 3a and 4a). When the model overprices an apartment, simulating the model well might actually anchor the participant on a value that is too high. In addition, since clear models present more information, participants in these conditions might be unlikely to notice unusual apartment configurations. Together, these factors could explain their high final predictions.

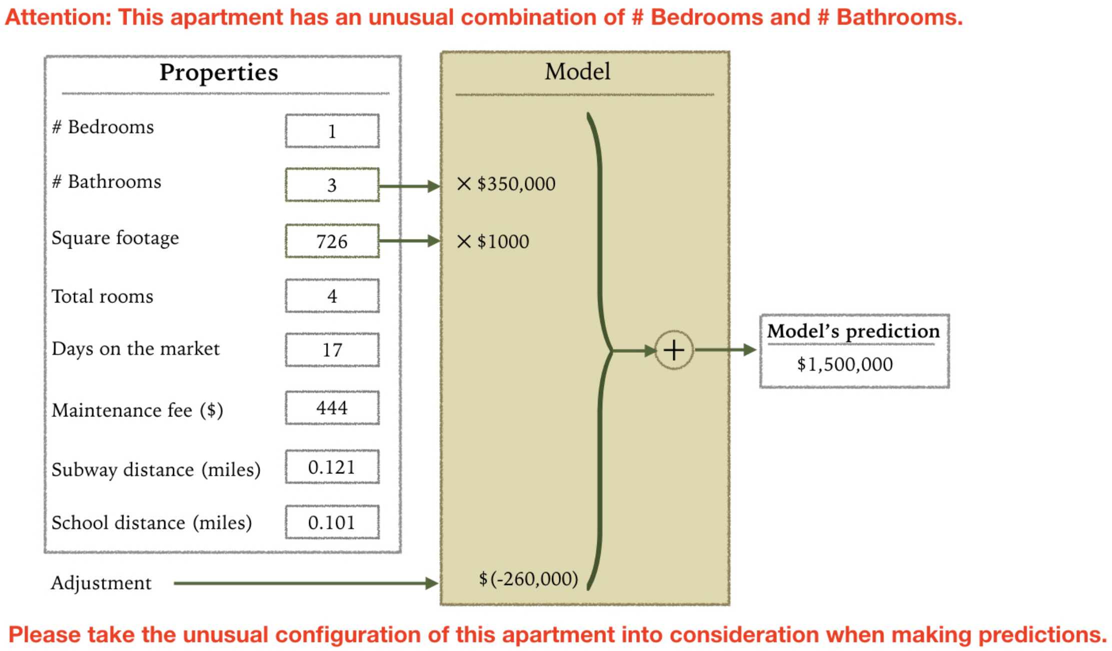
In contrast, participants in the black-box conditions could not simulate the model easily, but, undistracted by model internals, might be more likely to notice unusual apartment configurations. Interestingly, participants in these conditions apparently (incorrectly) assumed the model would also take the unusual configurations into account and thus reported simulated values that were too low (relative to the model). When making their final predictions, participants in the black-box conditions therefore could have had two factors working in their favor: they were not overwhelmed by the model internals and they were anchored on the low values they had previously stated. Both of these factors could lower their final estimate for the unusual apartments and thus increase their predictive accuracy.

In brief, being overwhelmed by the internals of the clear models and possible anchoring effects on prior estimates could together account for the observed differences in noticing unreasonable predictions in our first two experiments as opposed to our third experiment. To test this theory, our fourth experiment was designed to remove possible sources of anchoring and then measure the influence of an attention check that ensures that participants notice unusual configurations. Before running this experiment, we pre-registered three hypotheses:[7]

H11. **Effect of attention checks.** Participants who are shown an attention check will exhibit different abilities to correct the model's mistakes compared to those who are not provided with such an attention check.

H12. **Effect of model transparency without attention checks.** Participants who are not provided with an attention check on the unusual configurations of apartments will exhibit different abilities to correct the model's mistakes based on whether they are shown a clear model or a black-box model.

[7] Pre-registered hypotheses are available at `https://aspredicted.org/5xy8y.pdf`.



Attention: This apartment has an unusual combination of # Bedrooms and # Bathrooms.

Properties

# Bedrooms	1
# Bathrooms	3
Square footage	726
Total rooms	4
Days on the market	17
Maintenance fee ($)	444
Subway distance (miles)	0.121
School distance (miles)	0.101

Adjustment

Model

× $350,000

× $1000

$(-260,000)

Model's prediction

$1,500,000

Please take the unusual configuration of this apartment into consideration when making predictions.

Figure 7: Apartment 6 in the CLEAR-ATTENTION condition from our fourth experiment. Participants were provided with an explicit attention check (in red) on the unusual configurations of apartment 6 and apartment 8 in the testing phase.

H13. **Effect of model transparency with attention checks.** Participants who are provided with an attention check on the unusual configurations of apartments will exhibit different abilities to correct the model's mistakes based on whether they are shown a clear model or a black-box model.

5.1 Experimental design

Similar to our first two experiments, we asked people to predict apartment prices with the help of a model. We considered four experimental conditions in a 2 × 2 design:

- Participants were randomly assigned to see the model internals (CLEAR), or see the model as a black box (BB).
- Participants were randomly assigned to receive an explicit attention check concerning the unusual configurations of apartments (ATTENTION), or not (NO-ATTENTION).

Since we did not find any significant effect of number of features on people's abilities to detect the model's mistakes in our previous experiments, we only considered the 2-feature conditions. A screenshot of an apartment for which participants are provided with an attention check is shown in Figure 7.

We again ran the experiment on Amazon Mechanical Turk. We excluded people who had participated in our first three experiments, and recruited 800 new participants all of whom satisfied the selection criteria from our first three

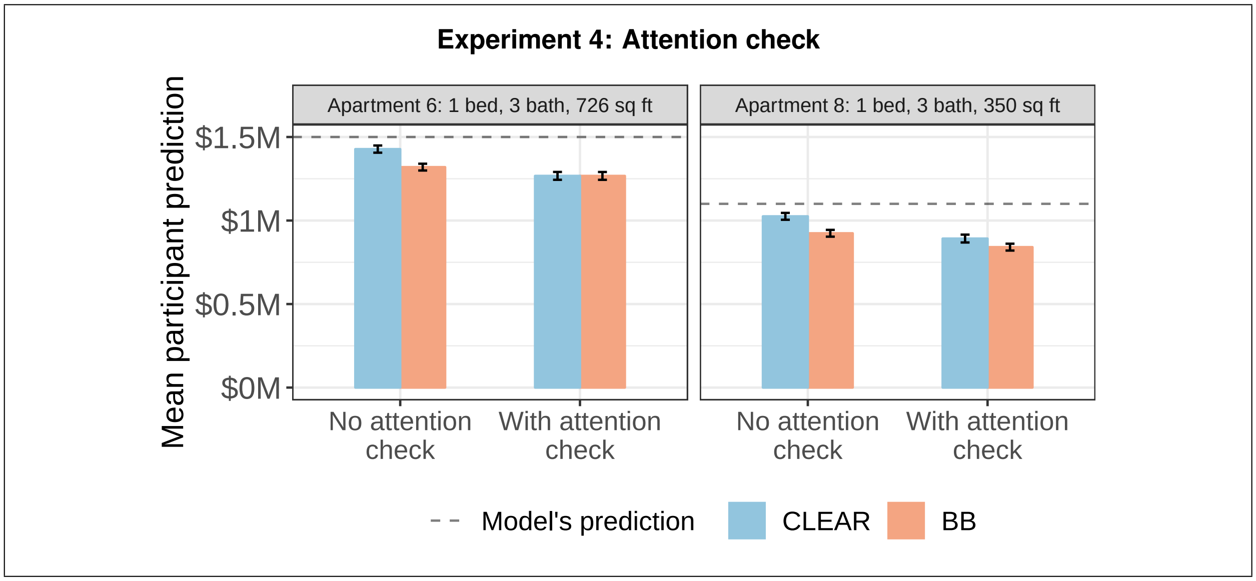


Figure 8: Mean predictions of the prices of apartments 6 and 8 in experiment 4. Error bars indicate one standard error.

experiments. The participants were randomly assigned to the four conditions (CLEAR-ATTENTION, $n = 202$; CLEAR-NO-ATTENTION, $n = 195$; BB-ATTENTION, $n = 201$; and BB-NO-ATTENTION, $n = 202$) and each participant received a flat payment of $\$1.00$.

We shortened the experiment by selecting five apartments from the training phase and five apartments from the first ten apartments in the testing phase of our previous experiments. The key items were three apartments which came after the first five apartments in the testing phase: an apartment from our dataset with normal configuration (one-bedroom, one-bathroom, 788 square feet) and two synthetically generated apartments with unusual configurations. The first synthetically generated apartment is apartment 12 from our previous experiments (one-bedroom, three-bathroom, 726 square feet, called *apartment 6*). The second synthetically generated apartment has a more unusual configuration (one-bedroom, three-bathroom, 350 square feet, called *apartment 8*), for which the model makes a unreasonably high prediction of the price. The order of the two synthetically generated apartments was randomized, while the apartment with normal configurations (*apartment 7*) was always shown in the middle.

Participants went through the exact same training phase as our first two experiments. For each apartment during the testing phase, participants were first shown the model's prediction and were asked to indicate how confident they were that the model was correct. Then, they were asked to make their own prediction of the apartment's price and to indicate how confident they were in this prediction.

5.2 Results

Figure 8 shows the mean deviation from the model's prediction for apartment 6 and apartment 8 from our fourth experiment.

H11. **Effect of attention checks.** Participants who were provided with an attention check deviated from the model's prediction more, on average, than participants who were not provided with an attention check for both apartment 6 ($t(791) = -4.72, p < .001$) and apartment 8 ($t(795) = -5.00, p < .001$).

H12. **Effect of model transparency without attention checks.** Of participants who were not provided with an attention check, those in the clear conditions deviated from the model's prediction less, on average, than participants in the black-box conditions for both apartment 6 ($t(393) = 3.65, p < .001$) and apartment 8 ($t(395) = 3.51, p < .001$).

H13. **Effect of model transparency with attention checks.** We found no significant difference in participants' deviation from the model between the CLEAR-ATTENTION and BB-ATTENTION conditions.

In line with our findings for the first five apartments, we found no difference in participants' deviation from the model's prediction for the normal apartment (7), suggesting there was not cognitive fatigue at the end of the task, as well as no effect of having seen an unusual configuration (the normal apartment was always presented between the unusual apartments).

In summary, we found that drawing attention to unusual apartment configurations in the absence of external anchors improved people's abilities to correct the model's mistakes to the extent that we no longer observe a significant difference between clear and black-box conditions on this measure. This result is consistent with our conjecture that the worse performance of participants in the clear conditions in detecting the model's mistakes is due to the overwhelming visual displays with information that draws participants' attention away from the peculiarities of the apartments.

6 Discussion and future work

We investigated how two factors that are thought to influence model interpretability—the number of features and model transparency—impact how well laypeople can simulate a model's predictions, how much they trust and follow those predictions, and how well they can correct unreasonable predictions. Although we found that a clear model with a small number of features was easier for participants to simulate, we found no difference in how closely people followed this model's predictions for typical examples as compared to a black-box model with many features. Even more surprisingly, we found that displaying model internals can have the unwanted effect of hampering people's ability to notice unusual inputs to a model and correct inaccurate predictions.

These findings have several implications for the design of interpretable model interfaces. First, when technically possible, it is helpful to highlight situations where a model receives unusual inputs that deviate from the data it was trained on. Second, it can be helpful to have people provide their own predictions before seeing the details or predictions of the model. Third, despite potential benefits of transparent models, it can in fact be detrimental to expose model internals by default, as this information can overwhelm end users. Instead, one might display only model inputs and predictions by default, hiding model internals until the user requests to see them.

This is not, however, to imply that transparency and parsimony should be ignored or avoided when designing machine learning models. Instead, it underscores the point that there are many possible goals in designing interpretable models, and that we should rely on rigorous user testing over intuition to assess whether those goals are met. Our experiments focused on just one type of model, presented to one subpopulation, for only a subset of the scenarios in which interpretability might matter.

For instance, it is likely the case that access to model internals allows people to better anticipate the inputs for which a model might go wrong. In fact, we leveraged this aspect of our own linear regression model for apartment prices to design the unusual configurations used in our experiments, since we could easily see that it would place unreasonably high value on additional bathrooms while keeping other features fixed. This almost certainly would have been more difficult with a black-box model. Likewise, it could be the case that people are more willing to adopt and deploy a clear model with few features compared to a more complex, black-box one.

Other possible extensions of our work include examining classification models (e.g., decision trees or rule lists) instead of regression models, along with different visualizations of these models. Our experiments could also be repeated with participants who are domain experts, data scientists, or researchers in lieu of laypeople recruited on Amazon Mechanical Turk. Other scenarios that could be explored include debugging a poorly performing model, assessing bias in a model's predictions, or explaining why an individual prediction was made. We hope this serves as a useful template for future research at the intersection of human-computer interaction and interpretable machine learning.

References

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. '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*, page 377. ACM, 2018.

[3] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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)*, 2015.

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

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

[6] Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 432. ACM, 2018.

[7] Francesca Gino and Don A. Moore. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1):21–35, 2007.

[8] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI)*, 2008.

[9] Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B Hamrick, and Patricia Chan. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 48(3):829–842, 2016.

[10] Jongbin Jung, Connor Concannon, Ravi Shro, Sharad Goel, and Daniel G. Goldstein. Simple rules for complex decisions. *arXiv preprint arXiv:1702.04690*, 2017.

[11] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[12] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *Proceedings of IEEE Conference and Visual Analytics Science and Technology*, 2017.

[13] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. *arXiv preprint arXiv:1805.11571*, 2018.

[14] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. In *FATML Workshop*, 2017.

[15] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2009.

[16] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[17] Jennifer M. Logg. Theory of machine: When do people rely on algorithms? Harvard Business School NOM Unit Working Paper No. 17-086, 2017.

[18] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[19] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

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

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

[22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 103. ACM, 2018.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[25] Richard Tomsett, David Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. In *2018 Workshop on Human Interpretability in Machine Learning*, 2018.

[26] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157): 1124–1131, 1974.

[27] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning Journal*, 102(3):349–391, 2016.

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

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