(De)Noise: Moderating the Inconsistency Between Human Decision-Makers

ANONYMOUS AUTHORS

Prior research in psychology has found that people's decisions are often inconsistent [53, 54]. An individual's decisions vary across time, and decisions vary even more across people. Inconsistencies have been identified not only in subjective matters, like matters of taste, but also in settings one might expect to be more objective, such as sentencing, job performance evaluations, or real estate appraisals. In our study, we explore whether algorithmic decision aids can be used to moderate the degree of inconsistency in human decision-making in the context of real estate appraisal. In a series of large-scale human-subject experiments, we study how different forms of algorithmic assistance influence the way that people review and update their estimates of real estate prices. We find that both (i) asking respondents to review their estimates in a series of algorithmically chosen *pairwise comparisons* and (ii) providing respondents with *algorithmic advice* are effective strategies for influencing human responses. Compared to simply reviewing initial estimates one by one, the aforementioned strategies lead to (i) a higher *propensity to update* initial estimates, (ii) a higher *accuracy* of post-review estimates, and (iii) a higher degree of *agreement* between the post-review estimates of different respondents.

CCS Concepts: • **Human-centered computing**; • **Information systems → Decision support systems**;

Additional Key Words and Phrases: Machine-Assisted Decision-Making, Algorithmic Decision Aids, Intelligent Decision Support Systems, Human Consistency, Inter-Annotator Agreement, Human-Centered AI

ACM Reference Format:
Anonymous Authors. 2023. (De)Noise: Moderating the Inconsistency Between Human Decision-Makers. 1, 1 (August 2023), 29 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

> "… humans are unreliable. Judgments can vary a great deal from one individual to the next, even when people are in the same role and supposedly following the same guidelines."
>
> Kahneman et al. [53]

Presented with identical information, the same person might make different decisions at different points in time, and the decisions of different people are likely to vary even more [53, 54]. Such inconsistencies between decision-makers have been identified in numerous settings including sentencing [6], job performance evaluations [89], real estate appraisals [2], and—especially close to the research community—conference reviewing [8, 10, 20, 59, 91].

In certain settings, variation in people's decisions is indispensable; it may contain invaluable information that reflects the variation in people's background knowledge, political or moral stances,

All authors contributed to conceiving the idea and designing the experiments. Anonymous Authors X, Y and Z contributed to interpreting the results and writing the paper. Anonymous Author X implemented the survey instruments and decision aids, conducted the exploratory analyses, and drafted Section 3.4 and the Appendix. Anonymous Author Y drafted the remainder of the paper, gathered the data, and conducted the confirmatory analyses.
Author's address: Anonymous Authors.

XXXX-XXXX/2023/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

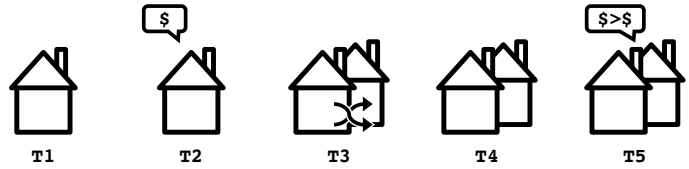


Fig. 1. Graphical overview of experimental conditions T1-T5. In T1 and T2, respondents review their decisions one-by-one, while in T3-T5 they review decisions in randomly (T3) or meaningfully selected (T4 and T5) pairs. In T2 and T5 respondents are additionally provided with (different kinds of) explicit machine advice.

life experiences, and other factors [84, 99, 100]. However, in other settings, consistency may be considered normatively desirable instead. For instance, Kahneman et al. [53] argue that organizations such as credit-rating and insurance agencies expect that, regardless of the particular professional handling each case, "identical cases should be treated similarly, if not identically"—a notion in line with that of "individual fairness" in the algorithmic fairness literature [24]. In conference reviewing, inconsistencies between different groups of reviewers have raised concerns about the peer-review process in the scientific community [8, 10, 20, 59, 91]. In the organizational justice literature, consistency of decisions is recognized as an important aspect of procedural justice [61, 63].

In this work, we focus on such settings where consistency between decision makers might be deemed desirable, and study how the degree of inconsistency of human decisions could be moderated with algorithmic assistance. This has immediate implications for the development of algorithmic assistance to support cooperative work by enabling the distribution of decision-making tasks amongst multiple decision-makers, without sacrificing consistency. Specifically, we leverage prior work in psychology and HCI to develop a set of algorithmic decision aids which may influence the degree of inconsistency of human decisions, and we rigorously experimentally evaluate the effect of these decision aids on human decisions.

Experimental Design. In a large-scale human-subject experiment ($N = 643$), we explored five different approaches to moderating human inconsistency, summarized in Figure 1. We focused on the task of estimating real estate prices, using a real estate price dataset studied by Poursabzi-Sangdeh et al. [78]. As a baseline (T1), we measured how people's decisions are affected by having the opportunity to review them one-by-one. We compared the effectiveness of this simple reviewing procedure with a series of more sophisticated interventions.

Inspired by the growing popularity of machine-assisted decision-making in the judge-advisor system (JAS) paradigm [9], we studied the effect of providing machine advice (T2). Leveraging prior research in psychology which found that people might find it easier to make pairwise comparisons than absolute judgments [73, 88], we also investigated the effectiveness of reviewing past judgments as a series of pairwise comparisons. We studied the effectiveness of asking respondents to review randomly selected pairs of apartments (T3), or pairs of apartments whose price estimates were inconsistent with most people's pairwise valuations—with (T5) and without (T4) explicitly informing participants about this inconsistency.

Contributions. We find that both providing traditional *machine advice* (T2) and asking respondents to review their past decisions as a series of meaningfully selected *pairwise comparisons* (T4 and T5) have a large and statistically significant effect on people's decisions.

Namely, in our pre-registered confirmatory analysis we find that compared to reviewing past decisions one-by-one (T1), our interventions (T2, T4 and T5) lead to:

- a higher *propensity of updating initial decisions*

- a higher *accuracy* of decisions after the review phase
- a higher degree of *consistency* amongst the post-review decisions of different respondents

In the paper's appendix, we additionally conduct an exploratory analysis of the effects of our interventions on different measures of accuracy and consistency.

Takeaways. Prior work on machine-assisted decision-making has showcased that traditional machine advice (T2) can influence people's decisions and—if the decision aid is sufficiently accurate—improve people's decision accuracy. We demonstrate that traditional machine advice can also increase the agreement between people's decisions.

We additionally go beyond studying traditional machine advice and demonstrate the effectiveness of novel alternative strategies. Asking people to review their decisions in a series of meaningfully selected pairwise comparisons (T4 and T5) significantly increases the accuracy and consistency of their decisions. Unlike traditional machine advice (T2), this approach does not require access to ground truth data—in our case, apartment prices. It requires only data about people's comparative valuations of apartments, which is used to identify pairs where a person's relative ordering of apartments does not align with the majority's estimates. Hence, it is also applicable in settings where ground truth may be costly to obtain or difficult to define.

Interestingly, when asking people to review meaningfully selected pairs of estimates (T4), we observe no significant effect of explicitly informing people that their relative ordering of apartments did not align with the majority's comparative valuations (T5). Hence, we identified an effective strategy for increasing the accuracy and consistency of people's responses that is applicable even in scenarios where it is normatively undesirable to explicitly steer people towards specific decisions.

Through our exploratory analysis in the appendix, we see that these results appear robust to alternative notions of accuracy and consistency.

2 BACKGROUND

Notions of Inconsistency. Much prior work has studied the (in)consistency of human decisions. A bulk of research has documented the *inconsistency between decisions of different decision-makers* [53, 54]—that is, a lack of *inter*-annotator consistency—in tasks as diverse as sentencing [6], evaluating job performance [89], estimating real estate prices [2], and reviewing conference submissions [8, 10, 20, 59, 91]. We contribute to this line of research by exploring if algorithms can be used to support cooperative work in such settings where consistency is deemed to be desirable. Namely, we propose methods for alleviating the inconsistency between the decisions of different decision-makers for the same set of inputs.

Much research has also studied the consistency of individual decision makers, or *intra*-annotator consistency. Cognitive biases such as dynamic inconsistency and hyperbolic discounting are known to result in the *inconsistency of an individual's decisions across time* [65, 90]. Individual's judgments are found to substantially vary across time in various settings [53]: pathologist's biopsy assessments of the same sample at different points in time were found to exhibit a correlation of only 0.61 [27]; expert's estimates of the amount of time required to complete the same software development task were found to vary by 71% [42].

Prior work has also documented the *inconsistency of an individual's judgments across inputs*. A particularly well-studied aspect of this problem is the inconsistency of people's pairwise preferences. Decades of research in this area have led to the development of numerous methods for identifying, measuring, and reducing the inconsistency of human pairwise preferences [1, 11, 55], as well as a plethora of approaches to the difficult task of learning human pairwise preferences [17, 33, 47], which we consult when developing our decision aids in Section 3.4.

Human Heuristics for Reducing Inconsistency. Many decision-making settings require people to make decisions on a case-by-case basis: granting or denying loans, making bail decisions, reviewing papers, etc. However, prior research in psychology has documented that people might find it easier to make *comparative* judgments than absolute ones in various contexts [73, 88]. Pairwise comparisons are also used to assist people with developing and refining their beliefs [62]. Hence, it is not surprising that people often rely on comparative judgments to assist them with making case-by-case decisions. For instance, the analytical hierarchy process that is widely used to assist with making complex decisions in domains ranging from governance to engineering relies on comparative judgments at its core [32, 93].

To illustrate how one may leverage comparative judgments to assist with absolute judgments, let us consider the task of grading papers. After (i) assigning initial grades to a set of papers, one might (ii) compare pairs (or larger subsets) of papers to identify mutually inconsistent decisions, in order to (iii) revise the final grades. In T4 and T5, we develop a decision aid that identifies pairs of decisions that are inconsistent with the majority's comparative valuations, hence providing a tool for automating step (ii) in the above-described procedure.

Reducing Inconsistency with Algorithmic Assistance. Much recent research has studied people's perceptions and utilization of algorithmic decision aids. Past research has compared how people react to human and machine advice [12, 22, 23, 25, 60, 66, 67, 70, 71, 79, 96], and studied the factors that influence the impact of machine advice [16, 28, 39, 40, 78, 86, 98, 101–103, 105]. The impact of machine advice was predominantly studied in terms of people's propensity to take machine advice [39, 101], and the effects of machine advice on the accuracy [40, 78, 105] and fairness [37, 38] of people's decisions. However, little prior work considered the effects of algorithmic assistance on the *consistency* of human decisions.

Kahneman at al. [53, 54] extensively study the problem of noise in human judgments. In Kahneman et al. [53], they discuss several approaches to reducing inconsistency in human decisions, proposing interventions of varying strengths. The first and most radical proposal is to replace human decision makers with algorithms. Still, they highlight the need for people to retain ultimate control. Hence, as the second and weaker proposal, they propose the use of algorithmic decision aids to assist human decisions. Depending on the estimated accuracy of human and algorithmic decisions and the normative importance of accuracy, they highlight the possibility of advising against overruling algorithmic predictions. The third and weakest intervention is ensuring that decision-makers use similar procedures to gather and integrate information, and to translate this information into a decision. The first two proposals rely on algorithmic assistance to reduce human inconsistency. The underlying idea is to use algorithms to *predict correct decisions*, which would steer or replace human decisions, thereby making them more consistent. In this paper, we also study an alternative approach: using algorithms to *identify inconsistencies* in human decisions, and to help people reduce their inconsistency by themselves.

Reactions to Feedback about Inconsistency. Algorithmic decision aids have proven to be effective in a plethora of settings. Here we review literature in social psychology that may help us form hypotheses about the effectiveness of algorithmic decision aids for the task of reducing inconsistency in human decisions, and guide the design of our decision aids. Specifically, we leverage prior work in social psychology to anticipate how people may react to being provided with feedback about their inconsistency.

Kahneman et al. [53] argue that inconsistency is undesirable in a variety of settings. If our respondents share this view, they might perceive feedback about their inconsistency as negative feedback. Prior work in organizational psychology has shown that people may not react positively to negative feedback about their performance. Negative feedback is not perceived as useful, results

in negative reactions, and is not associated with a recipient's willingness to change their behavior [87]. It is also found to evoke defensiveness and denial [68]. The main strategy employees use to reduce the impact of such negative feedback is to reject it [49]. To mitigate these effects, we will avoid framing the machine advice as negative feedback, and utilize strategies for softening the blow proposed by Steelman and Rutkowski [87]: providing high quality feedback delivered in a considerate manner.

Kahneman et al. [53] also show that people tend to vastly underestimate the degree of inconsistency in human decision-making. Hence, feedback about inconsistency may conflict with people's beliefs. Much prior work in psychology has found that people resist evidence that is contradictory to their preconceptions [5]. Two psychological concepts that are particularly relevant for predicting how people will react to conflicting information are cognitive dissonance [31, 44] and biased assimilation [69] or disconfirmation bias [26]. Both lines of research point to the same conclusion: due to overestimating their consistency, people may discount or reject the decision aids' feedback about their inconsistency. We attempt to mitigate this effect by familiarizing people with their lack of expertise with the task at hand: at the beginning of the experiment, participants complete a tutorial where they can observe the (in)accuracy of their real-estate price estimates.

3 METHODOLOGY

3.1 Experimental Design

In a large-scale, pre-registered human-subject experiment[1] run on Prolific we studied how different interventions influence people's estimates of real estate prices. The interventions included asking respondents to review their initial estimates one by one (treatments T1 and T2) or as pairwise comparisons (treatments T3, T4 and T5), and providing respondents with different forms of algorithmic assistance (treatments T2, T4 and T5).

Scenario. In this work, we focused on the task of estimating real estate prices. In our experiments, we utilized a dataset of New York City real estate prices introduced by Poursabzi-Sangdeh et al. [78]. The dataset contains information about 393 apartments located on the Upper West Side of New York City, which were listed for sale on the real estate website StreetEasy.com between 2013 and 2015. For each listing, we had access to basic information about the apartment, including the listing price, number of bedrooms, bathrooms and total number of rooms, the apartments' square footage and monthly maintenance fees, the number of days the apartment has been on the market, and the distance from the apartment to the nearest subway and school. We preprocess the data as proposed by Poursabzi-Sangdeh et al. [78]. Specifically, we remove the apartments where the number of bedrooms is greater than the total number of rooms or where the apartment's square footage is less than 200 sqft. With this preprocessing, we were left with 387 apartments. From these, we utilized 30 apartments as stimulus material in the human-subject experiments (Section 3.2), and the remaining 357 apartments for training the decision aids (Section 3.4).

Experimental Conditions. In each experimental condition, respondents were first asked to complete a tutorial, in order to familiarize themselves with real estate prices in New York City. Next, all respondents were asked to estimate the prices of the same 30 apartments. After gathering the respondents' initial estimates of apartment prices, we asked them to review their estimates in one of five different ways, as described below and summarized in Figure 1 and Table 1.

In the review phase, participants were randomly assigned to one of the following five conditions:

[1]Prior to conducting the human-subject experiment, we have obtained the approval of <our anonymous Institution's> ethical review board (ERB), and pre-registered our experiment on AsPredicted. The anonymized pre-registration documentation can be found on the following url: https://aspredicted.org/D7X_NKL.

	Reviewing Procedure	Algorithmic Assistance	Data Required
T1	one-by-one	none	none
T2	one-by-one	explicit advice	ground truth
T3	pairwise comparisons	none	none
T4	pairwise comparisons	implicit advice	human perceptions
T5	pairwise comparisons	implicit and explicit advice	human perceptions

Table 1. Overview of the characteristics of the 5 experimental conditions in our study. Reviewing Procedure: Are instances reviewed one-by-one or pairwise? Algorithmic Assistance: Do respondents have access to any form of algorithmic assistance? Data Required: Do the utilized decision aids require any type of labeled data?

T1: Respondents were asked to review all of their estimates *one-by-one*, in the same format as they originally made them: 30 apartments, one per page, shown in random order.

T2: Compared to T1, we manipulated the information provided in the review phase. Respondents were again asked to review all of their initial estimates one-by-one (30 apartments, one per page, shown in random order), but the apartment descriptions were accompanied by *machine advice*. Specifically, respondents were shown the estimates of a linear regression model which we trained to estimate real estate prices using the dataset introduced by Poursabzi-Sangdeh et al. [78], as described in Section 3.4. T2 corresponds to the standard machine-assisted decision-making setting in the judge-advisor system (JAS) paradigm [9].

T3: While T1 and T2 required respondents to review their decisions one-by-one, in T3–T5 decisions were reviewed in pairs. In T3, respondents were asked to review their decisions in a series of 15 *pairwise comparisons of randomly selected pairs* of apartments. Respondents were asked to review 15 pairs of apartments in order to keep the number of decisions reviewed equal to 30 across all treatments. The same apartment may have been shown in multiple pairs. Hence, even though the 15 pairwise comparisons provided respondents with 30 opportunities to update their estimates, this does not imply that they had an opportunity to update their initial estimate for each of the 30 unique apartments.

T4: We built upon T3 and attempted to meaningfully select the pairs of apartments to present. While T3 presents respondents with random pairs of apartments, T4 selects *pairs* where respondents' estimates are *not aligned with the majority's view*. Specifically, we implicitly provided machine assistance by asking participants to review pairs of apartments for which their initial price estimates did not align with most people's comparative valuations of those apartments. We trained a model to predict the majority's comparative valuations of apartment prices using a dataset of human-annotated pairwise comparisons of apartments that we gathered, as described in Section 3.4.

T5: The format of the review phase and the pair selection procedure remained the same as in T4, but we additionally explicitly informed people about the difference between their initial estimates and the predicted comparative valuations of most people.

Hypotheses. Building upon prior work on machine-assisted decision-making in psychology and HCI (reviewed in Section 2), we expect that the three forms of algorithmic assistance that we study will influence people's estimates, resulting in higher agreement across decision-makers, and—since the provided advice is more accurate than the average human's decisions[2]—higher accuracy as well. Specifically, we hypothesize that, compared with the control condition, our interventions T2, T4 and T5 will result in:

[2]More details about the decision aids' performance can be found in Section 3.4.

H1: A higher number of decisions updated in the review phase.
H1': A higher propensity to update decisions for the particular apartments shown in the review phase.
H2: A higher accuracy of post-review decisions.
H3: A higher degree of agreement between the post-review decisions of different respondents.

Since two of our algorithmic interventions (T4 and T5) rely on revising pairs of decisions, we include another treatment to explore the effectiveness of the pairwise approach without algorithmic assistance (T3). However, given our focus on algorithmic forms of assistance, and the lack of conclusive evidence about the benefits of pairwise comparisons in prior work, we do not hypothesize about the effects of T3, and only study its effects exploratively.

We include both hypotheses H1 and H1' since they capture different aspects of the interventions' effects on people's propensity to update decisions. H1 focuses on the overall effect of the intervention across all apartments, while H1' captures the effectiveness in prompting respondents to review their estimate for specific apartments that are shown in the review phase. (These are the same for T1 and T2.) The effect captured by H1 may be deemed more important in settings where the goal is to maximize the overall effect across all apartments, while H1' may be more appropriate if we are interested in measuring engagement with the algorithmic assistance.

Dependent Variables. For each apartment we measured the respondents' pre-review estimates and post-review estimates. Note that in treatments T3–T5, respondents were not given an opportunity to update their estimates for some apartments. In those cases, we defined their post-review estimate to be equal to their pre-review estimate.

To test our hypotheses, we formed dependent variables based on these measurements as follows:

H1 magnitude: Absolute difference between pre and post-review estimates.
H1' magnitude: Absolute difference between pre and post-review estimates, limited to only those apartments shown during the review phase.
H1 binary: 0 if the pre and post-review estimates are equal, otherwise 1.
H1' binary: 0 if the pre and post-review estimates are equal, otherwise 1, limited to only those apartments shown during the review phase.
H2: Difference between pre-review error and post-review error. The pre- and post-review errors are calculated as the absolute difference between the respondent's estimate and the ground truth for a given apartment.
H3: Difference between pre-review disagreement and post-review disagreement. The pre- and post-review disagreement are calculated as the absolute difference between the respondent's estimate and the average estimate (namely, the mean value of all respondents' estimates) for a given apartment.

Analysis. In Section 4, we report the findings of our confirmatory analyses related to hypotheses H1–H3. In the text, we report the findings of our statistical hypothesis testing, accompanied with plots that illustrate our findings using descriptive statistics. To test our hypotheses we rely on linear mixed models with crossed random effects for participants and apartments.[3,4,5] The dependent

[3]To alleviate convergence issues of models with crossed random effects, we initialize the starting values of the parameters to the estimated parameters of a simpler model—a linear mixed model with a random effects term for participants only.
[4]For hypotheses with binary dependent variables (H1 binary and H1' binary), we have also conducted the same analysis using a logistic regression. The results are qualitatively the same for both models. In the paper we report the results of the linear regression for ease of interpretation of the coefficients. For a discussion on the suitability of using linear models with binary dependent variables, we refer the readers to Hellevik [45].
[5]We additionally replicated all of our analyses using fixed effects models with two-way clustering of standard errors with respect to apartments and participants. To do so, we utilized the reghdfe Stata package [19] that implements the estimator described in Correia [18]. We found the results of both approaches to lead to consistent findings across all hypotheses.

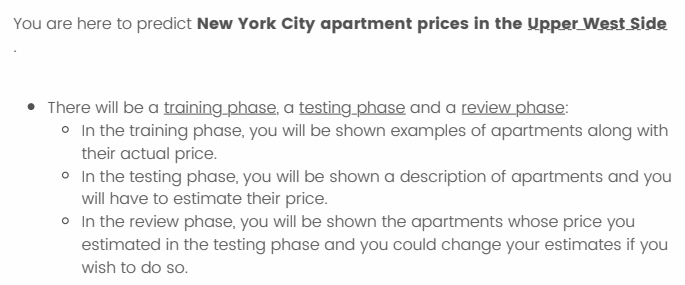


Fig. 2. Description of the experimental design shown to participants at the beginning of the experiment.

variables vary across hypotheses as described in the previous subsection. In all of the models, the experimental conditions are used as the independent variables, and treatment T1 is treated as the baseline. To compare the effects of other pairs of treatments we utilize Wald tests to test the equality of the corresponding coefficients. For the Wald tests, we report Bonferroni-adjusted p-values to account for the multiple comparisons problem.

In the appendix we conduct an additional exploratory analysis of our data. There we report a series of descriptive statistics related to the effect of the treatments on the respondents' accuracy and consistency.

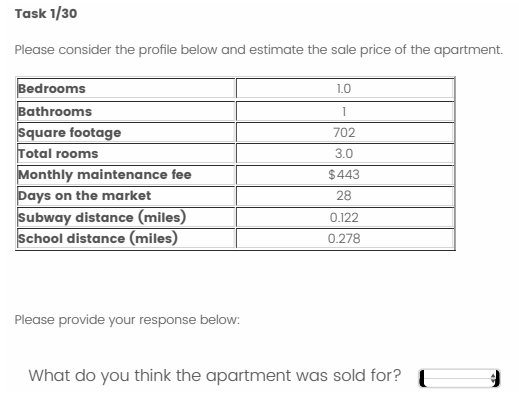
3.2 Stimulus Material

Upon opening the study link through the Prolific interface, participants were randomly assigned to one of the five experimental conditions. All participants first completed an online consent form and entered their Prolific worker ID. Next, participants were shown an introductory text describing the task (Figure 2).

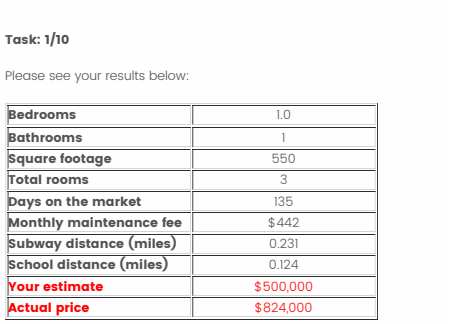
Following the approach of Poursabzi-Sangdeh et al. [78], participants were asked to complete a tutorial in order to familiarize themselves with real estate prices in New York City. The tutorial consisted of the same ten apartments that were utilized by Poursabzi-Sangdeh et al. [78]. The ten apartments were shown in random order, and for each apartment respondents were first asked to estimate its price based on its brief description (Figure 3a), and were then informed about the apartment's actual listing price (Figure 3b).

Next, we gathered the first part of our experimental data—the respondents' pre-review price estimates. We asked all participants to estimate the prices of the same 30 apartments. The 30 apartments were selected uniformly at random from the 387 apartments in the dataset, excluding the 10 apartments utilized in the tutorial. The set of apartments was kept constant across all experimental conditions and respondents. Throughout the experiment, in all five treatments, the apartments were shown in random order to avoid order bias [43, 83]. The phrasing and format of the questions and response options were identical to the tutorial (Figure 3a), except that we did not provide respondents with information about the apartments' true listing price after they reported their estimates.

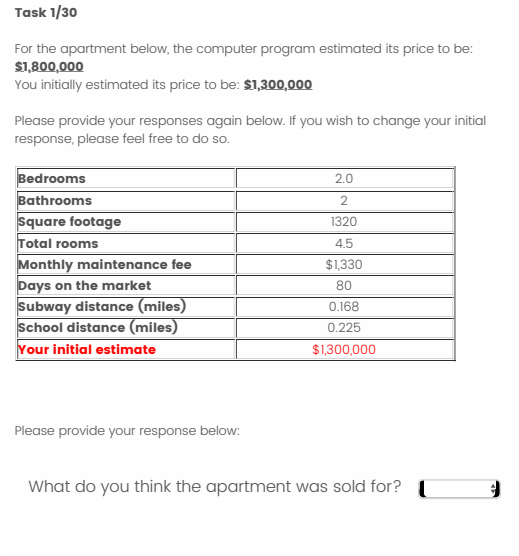
The respondents were then asked to respond to one simple instructed response item, which served as an attention-check question. Specifically, respondents were asked to "Please respond to this question by selecting Somewhat disagree as the answer", using a 5-point Likert scale as the response options. Similar instructed response items are commonly used for quality assurance purposes in online surveys, as a means of identifying inattentive or careless respondents [72].



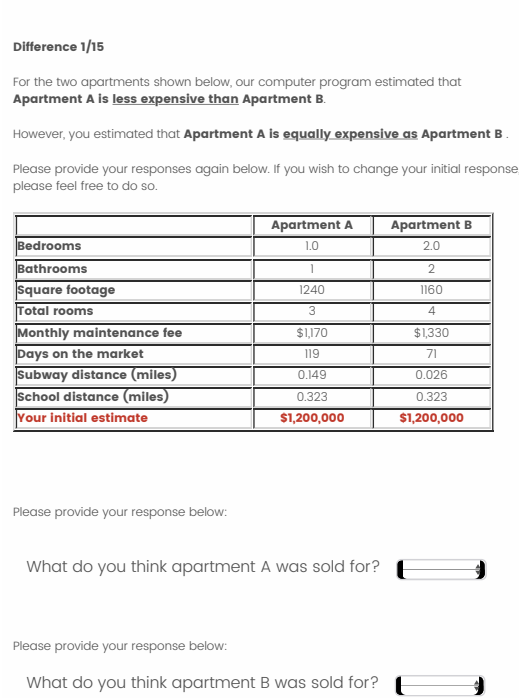
(a) Survey instrument used in the tutorial and for gathering respondents' pre-review estimates, across all treatments.



(b) Feedback provided to respondents during the tutorial, in all experimental conditions.



(c) Survey instrument used for gathering participants' post-review estimates in T2. In T1, the machine prediction was omitted.



(d) Survey instrument used for gathering participants' post-review estimates in T5. In T3 and T4, the machine prediction was omitted.

Fig. 3. Stimulus material

Next, we gathered the second part of our experimental data—the respondents' post-review price estimates. Respondents were presented with a text describing the experimental condition they were assigned to, i.e., they were informed about the procedure they will follow in the review phase. Participants were asked to review their initial responses in one of five different ways, based on

Demographic Attribute	Sample	Census
<35 years	56.4%	46%
35–54 years	32.6%	26%
55+ years	10.8%	28%
Male	50%	49%
Asian	10.3%	6%
Black	10.5%	12%
Hispanic	-	18%
Mixed	6.1%	-
White	68.5%	61%
Other	4.6%	4%

Table 2. Demographics of our study sample, compared to the 2019 U.S. Census [92].

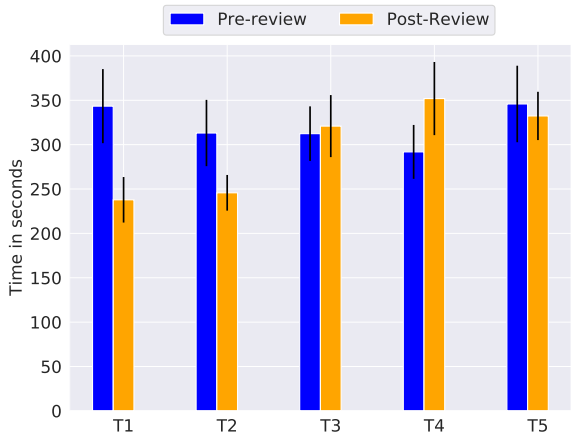


Fig. 4. Average duration of the experiment, per experimental condition, and per experimental phase. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM).

the experimental condition they were randomly assigned to. The experimental conditions T1–T5 are described in Section 3.1 "Experimental Conditions" and depicted in Figures 3c and 3d. All respondents reviewed 30 of their estimates. However, depending on the treatment, respondents reviewed all of their initial estimates one by one (T1 and T2) or a subset of their initial estimates in a series of pairwise comparisons with possible repetition of the apartments, up to 3 times, in multiple pairs (T3, T4 and T5).

Finally, we gathered participants' feedback about their experience of participating in the experiment. Namely, we asked respondents to tell us how much they agree with the following statements on a 5-point Likert scale from "Strongly agree" to "Strongly disagree": (i) The study was interesting, (ii) I would like to take part in a similar study in the future, (iii) The questions were easy to understand, (iv) The study was too long. At the end of the study the respondents also had the option to provide additional comments that they wanted to share with the researchers.

3.3 Data Collection

We recruited participants from Prolific—an online crowdsourcing platform which caters to scientific researchers [75]. Using Prolific's built-in pre-screening capabilities, we targeted respondents who: (i) are located in the US, (ii) have participated in at least 10 Prolific studies in the past, and (iii) have an approval rate of at least 95% on these past studies. We additionally utilized Prolific's option to provide a sample of respondents that is balanced with respect to gender, due to the current gender imbalance on the platform [14].

Our goal was to recruit sufficiently many participants to detect medium-sized effects (Cohen's $d = 0.5$) at the significance level of $\alpha = 0.05$ with power $\beta = 0.95$. Using the statistical software G*Power [29, 30], we calculated that a conservative Wilcoxon-Mann-Whitney two-tailed test requires 110 respondents per treatment group to detect effects of the size, significance level and power of interest. In our study, we have five experimental conditions, leading us to a minimum sample size of 550 respondents. To account for possible exclusions, incomplete or missing responses, we increased this estimate by 20% to 660 respondents.

We recruited a total of 660 participants from Prolific, over the course of several days (30th November – 3rd December 2022) in order to minimize sampling bias that could occur due to the day in the week or the time of day [13]. Participants were paid GBP 3.1 for taking part in the study. On average, participants were paid GBP 12.03 per hour, i.e., approximately USD $14.80 per hour—well above the federal minimum wage of USD $7.25. The median study completion time was 15 minutes and 28 seconds. The duration of the experiment varied across treatments, as depicted in Figure 4. As expected, there were no statistically significant differences between the average time taken to complete the pre-review phase of different experimental conditions. However, we observe significant differences across treatments in the review phase. Specifically, the review phase in T1 and T2—where respondents reviewed decisions one-by-one—took significantly less time than in T3–T5, where respondents reviewed pairs of decisions. When comparing the duration of the pre-review and review phase, T1 and T2 led to a significant increase in speed. On the other hand, the review phase in T4, where meaningfully selected pairs were presented without explicit advice, took more time compared to the pre-review phase. In T3 and T5 both the pre-review phase and the review phase took a similar amount of time to complete.

We report the demographics of our sample in Table 2. Since none of our hypotheses rely on demographic data, we did not ask our respondents to complete a demographics survey, in order to minimize the duration of our experiment, and to align with the data minimization principle. Hence, we report the data about our participants that we had access to through the crowdsourcing platform Prolific. Please note that this demographic data was self reported by Prolific crowdworkers directly to Prolific. Compared to the US census, our sample is younger, in line with typical samples recruited via online crowdsourcing platforms [46, 76, 85]. In line with the gender balancing pre-screening criteria employed during sampling, our sample is balanced with respect to gender. In terms of ethnicity, we are not able to directly compare our sample to the US census, since Prolific's simplified ethnicity prompt did not offer "Hispanic" as a response option. However, we note that Asian respondents are slightly over-represented and Black respondents are slightly underrepresented compared to the US census data.

Upon completing the study, participants were asked to provide feedback about their experience of taking part in this study. Most was positive. On a 5-point Likert scale from "Strongly agree" (coded as 5) to "Strongly disagree" (coded as 1), participants agreed with the statements "The study was interesting" ($\mu = 4.1 \pm 1.0$), "I would like to take part in a similar study in the future" ($\mu = 4.4 \pm 0.9$), and "The questions were easy to understand" ($\mu = 4.6 \pm 0.8$), while they neither agreed nor disagreed with the statement "The study was too long" ($\mu = 2.7 \pm 1.1$).

For the purposes of our analyses, we excluded all responses from participants who did not complete the full study (i.e., missing or incomplete responses), or who failed the instructed response attention check questions. A total of 17 respondents (2.6%) failed the attention check, leaving us with a final sample of 643 respondents.

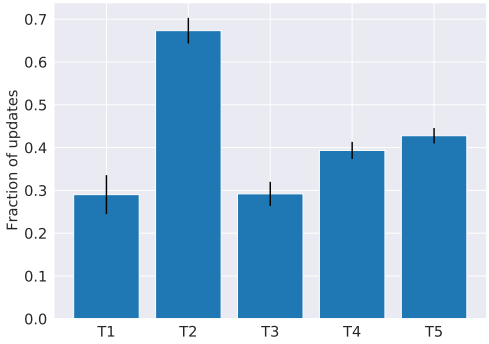
3.4 Decision Aids

Developing the Decision Aid Utilized in T2. In T2 we utilized a decision aid that *predicts apartment prices* to provide people with machine advice. Namely, we trained a linear regression model that used the apartment's attributes as independent variables (full list of features shown in Figure 3a), and the apartment price as the dependent variable. We normalized the independent variables to have zero mean and unit variance. We considered models with L1 (lasso) and L2 (ridge) regularization and without regularizers, and picked the regularization hyperparameter values which resulted in the highest coefficient of determination (R^2) on a 20% held out validation set. The model without any regularizer yielded the highest R^2 value: 85.6 on a test set comprised of 30% of the data. Amongst all of the features used in the final model, "Square footage" exhibited the strongest positive correlation with apartment price (with a weight of 1.08), while "Total rooms" exhibited the strongest negative correlation with price (with a weight of -0.3). In T2, we provided this model's estimates rounded to the nearest $100,000 as machine advice, for ease of interpretation.
Developing the Decision Aid Utilized in T4 and T5. For T4 and T5 we trained a decision aid to *predict people's comparative valuations of pairs of apartments*. We utilized this decision aid to identify people's estimates that do not align with the majority's pairwise comparisons.

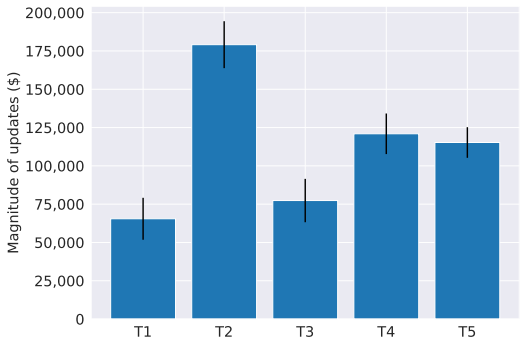
In order to build this tool, we gathered a dataset of human comparative valuations of apartments. We randomly selected 1000 unique pairs of apartments from our dataset and split them into 40 batches of 25 pairs. We recruited a total of 850 Prolific workers, who were randomly assigned to one of the batches. After excluding respondents who failed the attention-check question, we were left with 806 participants. From these, we excluded the last 6 responses so that each batch was labelled by exactly 20 participants. We gathered the data over several days (8th November – 11th November 2022) to minimize any bias caused by the time at which the data was gathered [13]. The participants were paid GBP 2 for taking part in the study, resulting in an average hourly rate of approximately USD $13.50.

The stimulus material and the experimental procedure were similar to the ones used in the main experiment, described in 3.2. The participants completed a consent form and entered their Prolific worker IDs, prior to observing an introductory text similar to the one shown in Figure 2. The participants then completed the same tutorial as in the main experiment, in which they were asked to estimate the prices of 10 apartments prior to observing their actual listing price, as shown in Figure 3b. Finally, respondents were asked to compare pairs of apartments, which were presented as shown in Figure 3d. Specifically, they were asked to estimate if "Apartment A" or "Apartment B" were more expensive. Additionally, they were asked how confident they were in their estimate on a 5 point Likert scale, ranging from "Completely guessing" to "Completely confident." To avoid order bias [43, 83], both the order of the 30 pairs of apartments and the order of apartments within a given pair were randomized.

Using this data, we trained a cross validated logistic regression classifier with L2 regularization to predict which apartment is perceived as more expensive in a given pair. To form the independent variables for our classifier, we subtracted the features of pairs of apartments (shown in Figure 3a) from one another. We then normalized them to have zero mean and unit variance. As the dependent variable we used the confidence weighted majority votes of the participant's responses. E.g., if a participant was "Completely guessing" their vote would count as $\frac{1}{5}$ and if they were "Completely



(a) H1 binary: fraction of decisions updated. The y-axis shows the fraction of the 30 initial decisions that were updated in the review phase.



(b) H1 magnitude: magnitude of updates. The y-axis shows by how much were the 30 initial decisions updated in the review phase.

Fig. 5. H1: Effect of the interventions on people's propensity to update decisions, across all 30 apartments. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM).

confident" it would count as 1. In the trained classifier "Square footage" was the most important feature, i.e., it had the largest absolute weight (36.18). The second most important feature was "Maintenance cost" (18.3). The average accuracy of our classifier on randomly chosen 30% test sets was 98.4%, and the accuracy of the classifier's predictions was positively correlated with the classifier's confidence.

In order to provide assistance to participants in T4 and T5, we used the aforementioned classifier. First, we converted the 30 apartments used in the main study into 435 pairs by taking all possible combinations, and predicted which of the apartments in each pair would be perceived as more expensive by most people. Then we ordered the pairs in a decreasing order with respect to the classifier's confidence. In T4 and T5 we iterated through this list, and asked participants to review their initial decisions which did not align with our classifier's predictions. E.g., if they initially estimated that Apartment A cost $600,000 and that Apartment B cost $900,000, while the classifier predicted that most people would perceive Apartment A as more expensive than Apartment B, participants could have been asked to review this pair of apartments. Participants were asked to review 15 pairs of apartments from this list. A single apartment was limited to appear in at most 3 pairs, to avoid negative reactions from repeatedly being presented with the same information. While selecting the pairs to show respondents, we took into account the decisions they may have updated during the review process.

4 RESULTS

In this section, we present the results of our analysis. We compare the baseline reviewing procedure T1 to our interventions T2, T4 and T5, in terms of their effect on people's propensity to *update* their initial estimates (H1 and H1'), and the *accuracy* (H2) and *consistency* (H3) of people's estimates.

4.1 H1: Overall Change in Decisions

In all five experimental conditions, we observe that people update some of their 30 initial decisions in the review phase. However, the number of decisions that are updated and the magnitude of these updates varies substantially between the experimental conditions. Compared to the control

	H1: change, bin., overall	**H1'**: change, bin., specific	**H1**: change, mag., overall	**H1'**: change, mag., specific	**H2**: accuracy	**H3**: agreement
T2	0.383***	0.383***	113634.1***	113634.1***	97299.5***	119916.8***
	(0.0216)	(0.0260)	(9608.2)	(13339.4)	(6022.1)	(6840.6)
T3	0.00179	0.157***	11897.4	52257.4***	-7141.5	-10044.1
	(0.0215)	(0.0264)	(9589.6)	(13489.2)	(6010.5)	(6827.3)
T4	0.103***	0.395***	55471.8***	140695.3***	25568.7***	24214.8***
	(0.0218)	(0.0269)	(9685.0)	(13713.6)	(6070.3)	(6895.3)
T5	0.138***	0.449***	49783.9***	130046.7***	20490.2***	24042.4***
	(0.0216)	(0.0266)	(9608.2)	(13596.1)	(6022.1)	(6840.6)
Cons.	0.290***	0.290***	65461.5***	65461.5***	7563.6	2271.4
	(0.0212)	(0.0216)	(11207.0)	(13277.9)	(6679.6)	(5555.7)
N	19290	14659	19290	14659	19290	19290

Table 3. Linear mixed models with crossed random effects for participants and apartments. The dependent variables for different hypotheses are described in Section 3.1. Experimental condition T1 is treated as the reference category in all models. I.e., intuitively, the row "Cons." shows the estimated value of the constant term (or intercept) that corresponds to the effects of treatment T1, while the rows T2-T5 show how the effects of these treatments differ compared to T1. Hence, to reason about the effects of T2-T5, one needs to sum up the values of the constant term and the treatment of interest. Standard errors are shown in parentheses. Statistical significance of coefficients is indicated as follows: * p < 0.05, ** p < 0.01, *** p < 0.001.

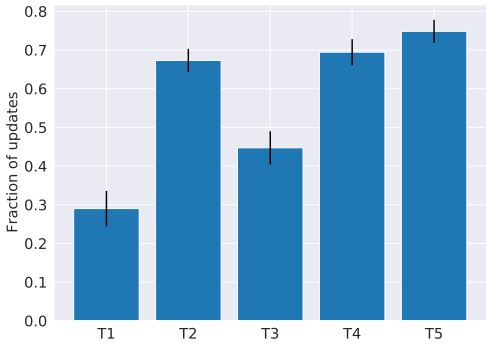
condition T1 our interventions T2, T4 and T5 lead to a higher propensity to update decisions in the review phase. That is, **our results support H1**. This holds both in terms of the number of decisions that were updated (H1 binary) and the magnitude of the change (H1 magnitude).

H1 binary: Number of Decisions Updated. Descriptively, we find that the fraction of decisions that are updated varies between treatments (first column of Table 3 and Figure 5a). In both T1 and T3 people update approximately 29% of their decisions, i.e., they update the estimated prices of 8.7 out of 30 apartments on average. In T4 and T5 people update a larger fraction of their decisions than in T1 and T3—close to 39% (11.8/30 apartments) and 43% (12.8/30 apartments) respectively. The treatment T2 has proven to be the most effective in prompting people to update their decisions, with approximately 67% of decisions (20.2/30 apartments) being updated.

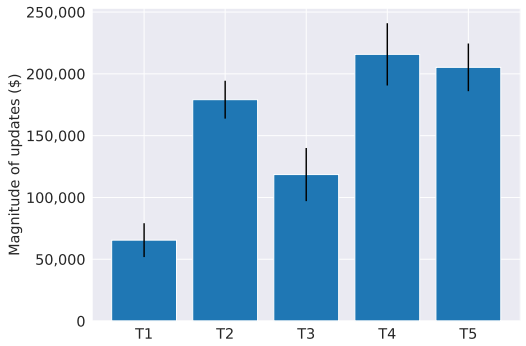
These descriptive observations are corroborated by our statistical analyses. The regression in the first column of Table 3 shows that all five treatments significantly influence human decisions. T2, T4 and T5 are significantly more effective than T1, while the effect of T3 was not significantly different than that of T1. Subsequent Wald tests performed on the estimated model confirmed that T4 and T5 are also more effective than T3 ($p < 0.001$), but did not identify a significant difference between the effects of T4 and T5 ($p = 0.35$). Finally, T2 was shown to be significantly more effective than all of the other treatments ($p < 0.001$).

That is, we find that people are more likely to update their decisions when reviewing meaningfully selected pairs of apartments and when machine advice is provided.

H1 magnitude: Magnitude of Updates. Our findings related to the magnitude of the changes are aligned with the findings about the number of decisions updated (third column of Table 3 and Figure 5b). In T1, people update their decisions by approximately $65,461 on average. In T3, the



(a) H1' binary: fraction of decisions updated. The y-axis shows the fraction of the decisions shown in the review phase that were updated.



(b) H1' magnitude: magnitude of updates. The y-axis shows by how much were the decisions shown in the review phase updated.

Fig. 6. H1': Effect of the interventions on people's propensity to update decisions, across the subset of apartments that were shown in the review phase. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM).

average update is close to $\$77,359$. The effect of both T1 and T3 is significantly different than zero ($p < 0.001$), and the difference between these two treatments is not statistically significant ($p = 0.214$). In T4, the average magnitude of the change was close to $\$120,933$. This is a significant increase compared to both T1 and T3 ($p < 0.001$). In T5, people updated their decisions by $\$115,245$ on average, which is significantly more than T1 and T3 ($p < 0.001$), but not significantly different than T4 ($p = 1$). Finally, people changed their decisions by close to $\$179,096$ in T2. The magnitude of this change is significantly larger than in any of the remaining treatments ($p < 0.001$).

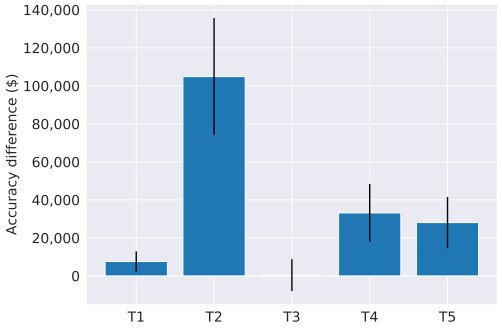
In short, we find that people update their decisions by a larger amount when they review them as a series of meaningfully chosen pairwise comparisons and when they observe machine advice.

4.2 H1': Propensity to Change Particular Decisions

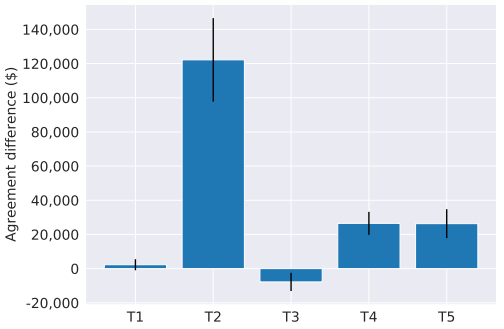
H1 considers the overall effect of our interventions across all 30 apartments. However, in T3–T5 participants were able to update only a subset of their initial decisions. In H1' we account for this and focus on the effect of our interventions across the apartments shown in the review phase.

In all five experimental conditions, respondents updated some of the decisions they were shown in the review phase. As in H1, the number of decisions that were updated and the magnitude of the updates varied significantly between treatments. When compared to the baseline treatment T1, our interventions T2, T4 and T5 result in a higher propensity to update decisions for the particular apartments shown in the review phase. I.e., **our findings support H1'**. Again, this holds both for the number of decisions that were updated (H1' binary) and the magnitude of change (H1' magnitude).

H1' binary: Number of Decisions Updated. The second column of Table 3 and Figure 6a provide information about the fraction of apartments shown in the review phase that respondents updated. For treatments T1 and T2 the results are identical to those related to H1, since all 30 apartments were shown in the review phase. Namely, in T1 respondents updated 29% of their decisions (8.7/30 apartments), while they updated 67% of their decisions (20.2/30 apartments) in T2. For T3–T5, results change substantially once we account for the fact that respondents could not update all 30 apartments in the review phase. While T3 was not significantly different than T1 in H1, in H1' we identified a significant difference between these two treatments. Namely, respondents updated 44%



(a) H2: effect on the accuracy of respondents' decisions. The y-axis shows the difference between the error in people's pre-review and post-review decisions.



(b) H3: effect on the agreement between respondents' decisions. The y-axis shows the difference between the disagreement in people's pre-review and post-review decisions.

Fig. 7. H2 an H3: Effect of the interventions on respondents' accuracy and inter-respondent agreement. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM).

of the decisions they were shown in the review phase in T3. The effects of T4 and T5 were even stronger, with respondents updating 69% and 74% of decisions they had access to in the review phase.

Our statistical analyses indicate that all treatments were significantly more effective than the baseline treatment T1 ($p < 0.001$). Treatments T4 and T5 were not significantly different from each other ($p = 0.1497$), but they were both more effective than T3 ($p < 0.001$). Unlike in H1, T2 was not the most effective treatment. While it was significantly more effective than T1 and T3 ($p < 0.001$), it was not significantly different from T4, and it was less effective than T5 ($p = 0.0421$).

On a high level, we found that people are more likely to update their decisions when asked to review them as a series of pairwise comparisons and when they are provided with machine advice.

H1' magnitude: Magnitude of Updates. The results of our analysis about the magnitude of changes are in line with our results about the amount of decisions that were updated (fourth column of Table 3 and Figure 6b). For T1 and T2, the results are the same as in H1: people update their decisions by approximately $65,461 in T1 and by $179,096 in T2. In T3–T5 the magnitude of the updates is significantly higher than in H1, with an average update close to $117,719 in T3, $206,157 in T4 and $195,508 in T5.

T1 is significantly less effective than the remaining four treatments ($p < 0.001$), and T3 is in turn significantly less effective than the remaining three treatments ($p < 0.001$), which are not significantly different between each other.

In other words, respondents updated their decisions by a larger amount when reviewing them as a series of pairwise comparisons and when they had access to machine advice.

4.3 H2: Accuracy of Respondents' Decisions

Next, we study the impact of our interventions on the quality of the decisions—the accuracy of people's estimates. We find that interventions T2, T4 and T5 significantly improve the accuracy of people's post-review decisions, compared to the baseline treatment T1. That is, **our results are in line with H2**.

In H1 and H1' we found that all five of our experimental conditions influenced people's decisions. However, not all of the reviewing procedures led to an increase in the accuracy of people's decisions.

As shown in the fifth column of Table 3 and in line with Figure 7a, the reviewing procedure utilized in T1 and T3 did not lead to a significant increase in the accuracy of people's post-review estimates, compared to their initial estimates. However, the remaining treatments did have a significant positive effect. Both T4 and T5 led to an increase in accuracy that is significantly higher than the one observed in T1 and T3 ($p < 0.001$). In T4 people's estimates of apartment prices improved by an average of \$33, 132, and in T5 by \$28, 054. The difference between T4 and T5 was not significant ($p = 1$). T2 led to a significantly higher improvement in accuracy ($p < 0.001$) than the remaining treatments—people's post review estimates were closer to the ground truth by an average of \$104, 863, compared to their initial estimates.

That is, while all treatments influenced people's decisions, not all of them led to an improvement in the accuracy of people's decisions. Only reviewing meaningfully selected pairs of apartments and having access to machine advice increased the accuracy of people's estimates.

4.4 H3: Agreement Between Respondents' Decisions

Finally, we investigate the effects of our interventions on the agreement between the decisions of different respondents. The patterns we identify are qualitatively similar to those related to the accuracy of people's decisions (H2). Namely, our interventions T2, T4 and T5 lead to a significantly higher increase in agreement between the post-review decisions of different respondents, compared to the control condition T1. That is, **the results support H3**.

As shown in the sixth column of Table 3 and in line with Figure 7b, treatments T1 and T3 do not lead to an increase in people's agreement, while T2, T4 and T5 do. Compared to the agreement between respondents' pre-review estimates, T4 and T5 increase respondents' post-review agreement by \$26, 486 and \$26, 314 respectively. The difference between T4 and T5 is not significant ($p = 1$), and the increase observed in both of these treatments is significantly higher than the effects observed T1 and T3 ($p < 0.001$). Treatment T2 increases the degree of agreement in people's post-review decisions by an average of \$122, 188, and this effect is significantly higher than the ones observed in any of the other treatments ($p < 0.001$).

On a high-level, we found that while all treatments influenced the respondents' decisions, some of them did not have an impact on the degree of agreement between the estimates of different respondents. Comparisons of meaningfully selected pairs of apartments and access to machine advice have yet again proven to be effective strategies.

5 DISCUSSION

5.1 Design Implications

T2 - Traditional Machine Advice. Reviewing past decisions with access to machine advice has proven to be more effective than doing so without machine advice, not only in terms of people's propensity to update their decisions, but also in terms of increasing the accuracy and consistency of their decisions. Given the recent research and real world applications that have demonstrated that people are willing to take machine advice, particularly when the advice is highly accurate [101] (as is the case in the real estate appraisal setting we consider), this is hardly a surprise.

Our results indicate that in settings where (i) one has access to ground truth data that enables the development of accurate decision aids, and (ii) it is deemed normatively desirable or acceptable to explicitly steer people towards making decisions in line with machine predictions, traditional algorithmic decision aids are an effective tool for doing so.

T3 - Randomly Selected Pairwise Comparisons. Past research in psychology has found that people are better at making comparative judgments than absolute ones in certain contexts [73, 88]. Still, people's pairwise preferences are known to be inconsistent [1, 11, 55]. Building upon both

lines of research, we investigated if people's decisions may benefit from being reviewed in a series of pairwise comparisons, instead of one-by-one.

If this intervention had proven to be effective, it would have important design implications. This intervention would be suitable for low-resource environments, where it is difficult or impossible to develop machine decision aids, due to a lack of data for training them, the inherent difficulty of making accurate predictions in the decision-making task at hand, or a lack of well-established notions of objective ground truth.

However, in our experiments, we did not find a significant difference between the accuracy and consistency of decisions that were reviewed one-by-one and those reviewed in pairs. Hence, our results suggest that it might not be sufficient to switch from absolute to comparative decision-making when reviewing decisions.

T4 - Agreement-Based Pairwise Comparisons. We found that switching from an absolute to a comparative reviewing procedure is an effective strategy when respondents compare *meaningfully* selected pairs of inputs. In T3, respondents are simply asked to review random pairs of decisions, but in T4 the reviewing procedure is guided by a machine decision aid. At first, this may not be evident—traditional decision aids (such as the one utilized in T2) usually attempt to predict correct decisions, and provide people with those predictions as advice. The decision aid used in T4 is quite different—it attempts to predict typical human decisions, and asks people to review their decisions when they do not match the predicted ones. In doing so, this procedure might be identifying precisely the pairs that would prompt people to review their decisions.

While the decision aid utilized in T4 is less effective than the one from T2, it has two important advantages: (i) it does not require access to ground truth data, and (ii) it does not provide explicit advice on how to update decisions. The former makes this approach suitable even in environments where one does not have access to ground truth data, or when it is not possible to define it. The latter makes it applicable even in settings where it is not normatively desirable to explicitly steer people towards specific decisions (e.g., due to concerns about silencing minority opinions), but to prompt people to review their own decisions and make them mutually consistent. I.e., this intervention may be an excellent candidate for future research on *intra*-annotator notions of consistency.

T5 - Agreement-Based Pairwise Comparisons with Advice. Perhaps surprisingly, T5 was not significantly more effective than T4 with respect to any of the dependent variables we studied. Reviewing past decisions as a series of meaningfully selected pairwise comparisons is equally effective with and without explicit machine advice. Hence, when it is normatively undesirable to explicitly steer people towards specific response options, one can omit machine advice without impeding the positive effects of the reviewing procedure on the accuracy and consistency of human decisions.

In future research, it would be interesting to study why explicit machine advice does not have an effect in this setting. We hypothesize this might be caused by its redundancy: when comparing two apartments, respondents might be able to infer the majority's comparative valuation of these apartments. Research on incentive mechanisms that rely on people's ability to predict others' responses provides some backing to this hypothesis. Namely, peer prediction mechanisms [74], in particular Bayesian Truth Serums and similar methods [56, 80–82], ask respondents to predict what others will report in order to design proper incentives that incentivize truthful reporting. If people are able to accurately predict the majority's comparative valuations, explicit machine advice may not provide any additional information to respondents, and hence have no effect on their decisions. Future studies can test this hypothesis by evaluating people's ability to predict others' pairwise comparisons.

5.2 Limitations and Future Work

Notions of Consistency. In this paper, we explored the effects of our interventions on several measures of *inter*-annotator consistency. In future work, it would be interesting to go beyond inter-annotator consistency, and consider notions of *intra*-annotator consistency.

The simplest extension would be the study of intra-annotator consistency *across time*. That is, instead of measuring the degree of agreement between different respondents for the same input, one could measure the degree of consistency of the same annotator for the same input in different points in time. This extension requires minimal changes to our experimental design—namely, it requires conducting a longitudinal human-subject study. This line of work could provide important insights about moderating the effects of cognitive biases that lead to a person's inconsistency through time, such as dynamic inconsistency and hyperbolic discounting [65, 90], or the "hungry judge" effect [21].[6]

Intra-annotator consistency across time is closely related to counterfactual questions such as "Would the decision-maker have made the same decision in a different point in time?" A different notion of consistency—intra-annotator consistency *across inputs*—addresses the question "Does the decision-maker make similar decisions for similar inputs?" This line of research is closely related to research on individual fairness.

A central problem in studying both intra-annotator consistency across inputs and individual fairness lies in defining the similarity metric which determines which inputs should be treated as similar. Prior work on individual fairness has assumed such similarity metrics to be given [24], or defined them based on the inputs' ground truth labels [51, 64], distance in transformed feature spaces that align with certain distributive fairness criteria [57, 104], or—as we implicitly did in T4 and T5—based on human judgments about input similarity [50, 52, 58, 97]. As a promising direction for future work, we highlight the study of intra-annotator consistency across inputs with personalized similarity metrics, i.e., the development of methods for identifying decisions that are outliers, inconsistent with the other decisions made by the same respondent.

In this work we study methods for alleviating inter-annotator inconsistency. Deciding which notion of inconsistency is appropriate to apply in a given setting is inherently a normative question. Hence, we invite future work not only on formalizing and operationalizing different notions of inconsistency, but also philosophical and policy discussions on the desirability of different—and as discussed below, any—notions of consistency in specific settings.

Benefits of Human Inconsistency. This paper focused on settings where inconsistency between multiple decision-makers may be deemed undesirable. As such, the proposed methods are not applicable and should not be applied in settings where diversity in people's beliefs, perceptions, and behavior may be beneficial, or considered normatively desirable.

Diversity in people's decisions may reflect the differences in their skill set and background knowledge, and these differences can be exploited to improve decision-making quality [100]. Diversity in the composition of groups increases the diversity in the problem solutions that team members propose, which in turn increases the quality of group decisions [99]. Heterogeneity in teams can benefit group performance, since the diversity in the perspectives of different team members can foster creativity and innovation [84].

Furthermore, people's beliefs, perceptions and behavior are known to correlate with their socio-demographics and life experiences [3, 4, 34, 36, 41, 48, 77, 95]. That is, the decisions made by members of minority groups may systematically differ from those made by members of the majority. Therefore,

[6]While the "hungry judge" effect [21] is often referenced as an argument in favor of introducing algorithmic assistance in legal decision making, the validity of the study's findings has been much debated in recent literature [15, 35].

methods for reducing inconsistency between people may—inadvertently or on purpose—explicitly steer people's decisions toward the majority's view, thereby silencing the minorities' views.

Hence, prior to applying any methods for reducing disagreement between decision-makers, it is crucial to evaluate whether such an outcome would be appropriate and normatively desirable in the decision-making task at hand.

Generalizability to Other Domains. In this work, we focus on a real estate appraisal scenario. While we find large and statistically significant effects of our interventions for the task at hand, we invite future work that will systematically explore which types of scenarios our findings generalize to. We opted for this scenario because many laypeople have prior experience with property valuation (e.g., searching for, purchasing or selling real estate), but most laypeople do not make highly accurate estimates of real estate prices. The task we considered may be in the sweet spot between too difficult and too easy for our respondent sample. We hypothesize that our findings may not generalize to tasks on either of the extremes.

For tasks that people find easy, such as visual recognition tasks, interventions may not have an effect if people already exhibit high degrees of accuracy and agreement, hence not allowing room for significant improvement along either dimension. For tasks that people find difficult, such as criminal risk prediction, both people and algorithms may exhibit low levels of accuracy. For instance, in a pilot study we conducted using the ProPublica COMPAS dataset [7], algorithmic advice (T2, with an accuracy of 58%) did not have a significant impact on agreement since it increased agreement for some cases, while decreasing it for others. The latter typically occurred for the non-negligible number of cases where respondents initially made correct predictions, but incorrect machine advice steered them away from their initial responses, decreasing their accuracy and agreement levels.

Interventions. In this paper, we report the effects of five different interventions. In pilot studies we considered one additional intervention, where we asked respondents to review all of their initial estimates on the same page, sorted by the apartment prices they estimated. We initially conjectured that this may allow respondents to conduct comparisons of apartments that they deemed to have similar prices. However, since (i) this approach was not scalable to a large number of decisions, and (ii) the effects of this treatment showed no statistically significant difference from T1 and T3 in our pilots, we omitted it from our main study for brevity.

We invite future work that would explore an even broader set of interventions. As reviewed in Section 2, prior work on human advice taking behavior [9] and on machine-assisted decision-making has identified numerous factors that influence how people take advice, including the decision aid's accuracy [101], explainability [78], and the stakes associated with the decision-making task [39], and future work could incorporate some of these factors in their interventions. Future work could also build upon T4 and T5 by developing decision aids that not only predict which of two apartments is perceived as more expensive, but also identify apartments that are perceived to be equally expensive. Identifying data points that are perceived as deserving of similar outputs may be interesting not only for the study of noise in human decisions, but also for research on individual fairness.

Respondent Samples. In our experiments we recruited a large and demographically diverse set of laypeople from the US. Future work could explore if our findings replicate in other cultures beyond the US. Additionally, it is worth noting that our sample consisted of laypeople, and it is possible that the decisions of professionals such as real estate agents systematically differ from our lay sample. For instance, professionals may be substantially more accurate in their predictions, thereby having fewer opportunities to benefit from algorithmic advice. Hence, it may be interesting to replicate our experiments with industry professionals.

5.3 Conclusion

In this work, we studied methods for alleviating inconsistency in human decision-making. We identified several approaches that effectively influence human decisions, improving their accuracy and consistency with other respondents. We identified methods that are applicable to a wide variety of scenarios, including for settings where one has access to ground truth data for training decision aids (T2), as well as for settings where one only has access to human annotations (T4 and T5), but none for settings where no data is available (T3). All of the treatments that significantly improved decision accuracy and consistency relied on algorithmic assistance, be it explicit (T2 and T5) or implicit (T4). As a promising avenue for future work, we see the study of a broader set of notions of inconsistency, including intra-annotator consistency.

REFERENCES

[1] Edward Abel, Ludmil Mikhailov, and John Keane. Inconsistency reduction in decision making via multi-objective optimisation. *European Journal of Operational Research*, 267(1):212–226, 2018.

[2] Alastair Adair, Norman Hutchison, Bryan MacGregor, Stanley McGreal, and Nanda Nanthakumaran. An analysis of valuation variation in the uk commercial property market: Hager and lord revisited. *Journal of property valuation and Investment*, 1996.

[3] Michele Albach and James R Wright. The role of accuracy in algorithmic process fairness across multiple domains. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 29–49, 2021.

[4] Alberto Alesina and Paola Giuliano. Preferences for redistribution. In *Handbook of social economics*, volume 1, pages 93–131. Elsevier, 2011.

[5] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. *The nature of prejudice*. Addison-wesley Reading, MA, 1954.

[6] James M Anderson, Jeffrey R Kling, and Kate Stith. Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines. *The Journal of Law and Economics*, 42(S1):271–308, 1999.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016. Accessed: 2022-09-14.

[8] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Has the machine learning review process become more arbitrary as the field has grown? the NeurIPS 2021 consistency experiment. *arXiv preprint 2306.03262*, 2023.

[9] Silvia Bonaccio and Reeshad S Dalal. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2):127–151, 2006.

[10] Amy S Bruckman, Casey Fiesler, Jeff Hancock, and Cosmin Munteanu. CSCW research ethics town hall: Working towards community norms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pages 113–115, 2017.

[11] Matteo Brunelli and Michele Fedrizzi. Axiomatic properties of inconsistency indices for pairwise comparisons. *Journal of the Operational Research Society*, 66(1):1–15, 2015.

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

[13] Logan S Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z Strolovitch. Intertemporal differences among mturk workers: Time-based sample variations and implications for online data collection. *Sage Open*, 7(2): 2158244017712774, 2017.

[14] Nick Charalambides. We recently went viral on TikTok - here's what we learned. https://www.prolific.co/blog/we-recently-went-viral-on-tiktok-heres-what-we-learned, 2021. Accessed: 2022-09-14.

[15] Konstantin Chatziathanasiou. Beware the lure of narratives:"hungry judges" should not motivate the use of "artificial intelligence" in law. *German Law Journal*, 23(4):452–464, 2022.

[16] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW), 2023 (to appear).

[17] Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. Predicting partial orders: ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 215–230. Springer, 2010.

[18] Sergio Correia. A feasible estimator for linear models with multi-way fixed effects. Technical report, Duke University, 2016. Working Paper.

[19] Sergio Correia. reghdfe: Stata module for linear and instrumental-variable/gmm regression absorbing multiple levels of fixed effects. *Statistical Software Components s457874, Boston College Department of Economics*, 2017.

[20] Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.

[21] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

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

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

[24] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[25] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1):79–94, 2002.

[26] Kari Edwards and Edward E Smith. A disconfirmation bias in the evaluation of arguments. *Journal of personality and social psychology*, 71(1):5, 1996.

[27] Hillel J Einhorn. Expert judgment: Some necessary conditions and an example. *Journal of applied psychology*, 59(5): 562, 1974.

[28] Christoph Engel and Nina Grgić-Hlača. Machine advice with a warning about machine limitations: Experimentally testing the solution mandated by the wisconsin supreme court. *Journal of Legal Analysis*, 13(1):284–340, 2021.

[29] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.

[30] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

[31] Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1962.

[32] Ernest H Forman and Saul I Gass. The analytic hierarchy process—an exposition. *Operations research*, 49(4):469–486, 2001.

[33] Johannes Fürnkranz and Eyke Hüllermeier. Pairwise preference learning and ranking. In *Machine Learning: ECML 2003: 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 14*, pages 145–156. Springer, 2003.

[34] Paola Giuliano and Antonio Spilimbergo. Growing up in a recession. *The Review of Economic Studies*, pages 787–817, 2014.

[35] Andreas Glöckner. The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision making*, 11(6):601–610, 2016.

[36] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366, 2011.

[37] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *FAT**, 2019.

[38] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[39] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[40] Nina Grgić-Hlača, Claude Castelluccia, and Krishna P Gummadi. Taking advice from (dis) similar machines: The impact of human-machine similarity on machine-assisted decision-making. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 74–88, 2022.

[41] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M Redmiles. Dimensions of diversity in human perceptions of algorithmic fairness. *Proceedings of The second ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2022.

[42] Stein Grimstad and Magne Jørgensen. Inconsistency of expert judgment-based estimates of software development effort. *Journal of Systems and Software*, 80(11):1770–1777, 2007.

[43] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*. John Wiley & Sons, 2011.

[44] Eddie Harmon-Jones and Judson Mills. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In *Cognitive Dissonance, Second Edition: Reexamining a Pivotal Theory in Psychology*. American Psychological Association, 2019.

[45] Ottar Hellevik. Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, 43: 59–74, 2009.

[46] Connor Huff and Dustin Tingley. "who are these people?" evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics*, 2(3):2053168015604648, 2015.

[47] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17):1897–1916, 2008.

[48] Jon Hurwitz and Mark Peffley. Explaining the great racial divide: Perceptions of fairness in the us criminal justice system. *The journal of politics*, 67(3):762–783, 2005.

[49] Daniel R Ilgen, Cynthia D Fisher, and M Susan Taylor. Consequences of individual feedback on behavior in organizations. *Journal of applied psychology*, 64(4):349, 1979.

[50] Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

[51] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

[52] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. An algorithmic framework for fairness elicitation. *arXiv preprint arXiv:1905.10660*, 2019.

[53] Daniel Kahneman, Andrew M Rosenfield, Linnea Gandhi, and Tom Blaser. Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, 2016.

[54] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. *Noise: A Flaw in Human Judgment*. Little, Brown Spark, 2021.

[55] Waldemar W Koczkodaj. A new definition of consistency of pairwise comparisons. *Mathematical and computer modelling*, 18(7):79–84, 1993.

[56] Erin L Krupka and Roberto A Weber. Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524, 2013.

[57] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, pages 1334–1345. IEEE, 2019.

[58] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439*, 2019.

[59] Neil D. Lawrence. A retrospective on the 2014 NeurIPS experiment. http://inverseprobability.com/talks/notes/the-neurips-experiment.html, June 2021. Accessed: 2022-09-14.

[60] Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684, 2018.

[61] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[62] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35, 2019.

[63] Gerald S Leventhal. What should be done with equity theory? new approaches to the study of fairness in social relationships. In *Social exchange: Advances in theory and research*, pages 27–55. Springer, 1980.

[64] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

[65] George Loewenstein and Drazen Prelec. Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2):573–597, 1992.

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

[67] Jennifer Marie Logg. Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086*, 2017.

[68] Manuel London. *Job feedback: Giving, seeking, and using feedback for performance improvement*. Psychology Press, 2003.

[69] Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

[70] Poornima Madhavan and Douglas A Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, 2007.

[71] Hasan Mahmud, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390, 2022.

[72] Adam W Meade and S Bartholomew Craig. Identifying careless responses in survey data. *Psychological methods*, 17 (3):437, 2012.

[73] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[74] Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

[75] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[76] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 2010.

[77] Emma Pierson. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*, 2017.

[78] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

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

[80] Drazen Prelec. A bayesian truth serum for subjective data. *science*, 306(5695):462–466, 2004.

[81] Goran Radanovic and Boi Faltings. A robust bayesian truth serum for non-binary signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 833–839, 2013.

[82] Goran Radanovic and Boi Faltings. Incentives for truthful information elicitation of continuous signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[83] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical report, University of Maryland, 2017.

[84] Marie-Élène Roberge and Rolf Van Dick. Recognizing the benefits of diversity: When and how does diversity increase group performance? *Human Resource management review*, 20(4):295–308, 2010.

[85] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. Who Are The Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI*, 2010.

[86] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1–8. IEEE, 2015.

[87] Lisa A Steelman and Kelly A Rutkowski. Moderators of employee reactions to negative feedback. *Journal of Managerial Psychology*, 19(1):6–18, 2004.

[88] Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

[89] Robert L Taylor and William D Wilsted. Capturing judgment policies: A field study of performance appraisal. *Academy of Management Journal*, 17(3):440–449, 1974.

[90] Richard Thaler. Some empirical evidence on dynamic inconsistency. *Economics letters*, 8(3):201–207, 1981.

[91] David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2020.

[92] U.S. Census Bureau. American Community Survey 5-Year Estimates, 2019.

[93] Omkarprasad S Vaidya and Sushil Kumar. Analytic hierarchy process: An overview of applications. *European Journal of operational research*, 169(1):1–29, 2006.

[94] Raphael Vallat. Pingouin: statistics in python. *The Journal of Open Source Software*, 3(31):1026, November 2018.

[95] Florian Van Leeuwen, Bryan L Koenig, Jesse Graham, and Justin H Park. Moral concerns across the united states: Associations with life-history variables, pathogen prevalence, urbanization, cognitive ability, and social class. *Evolution and Human Behavior*, 35(6):464–471, 2014.

[96] Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. *arXiv preprint arXiv:2107.07015*, 2021.

[97] Hanchen Wang, Nina Grgic-Hlaca, Preethi Lahoti, Krishna P Gummadi, and Adrian Weller. An empirical study on learning fairness metrics for compas data with human supervision. *arXiv preprint arXiv:1910.10255*, 2019.

[98] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[99] John P Wanous and Margaret A Youtz. Solution diversity and the quality of groups decisions. *Academy of Management journal*, 29(1):149–159, 1986.

[100] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.

[101] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[102] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 223–227, 2016.

[103] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 307–317, 2017.

[104] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[105] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *arXiv preprint arXiv:2001.02114*, 2020.

A APPENDIX

A.1 Other Measures of Accuracy: an Exploratory Analysis

In Section 4, we studied the effect of our interventions on the accuracy of peoples' estimates of apartment prices. In this section, we go beyond the accuracy of people's absolute judgments about apartment prices, and consider the accuracy of their implicit relative judgments.

We start by deriving people's implicit relative judgments from their absolute estimates. For each pair of apartments (A,B), we check if a respondent estimated Apartment A to be more expensive (>), less expensive (<) or equally as expensive (=) as Apartment B. Then we compare these implicit relative judgments with the ground truth (i.e., with the relative ordering of apartments based on their listing price).

In Figure 8, we report the fraction of instances where people's implicit relative ordering differed from the apartments' true ordering. Descriptively, we observe that the error in people's pre-review relative estimates is similar across all experimental conditions. In experimental conditions T1 and T3, the error in people's pre- and post-review estimates remained similar. However, in T2, T4 and T5, the error decreased by 5.2, 5.1 and 5.0 percentage points respectively. That is, treatments T2, T4 and T5 reduced the error in people's implicit relative judgments by 22.1%, 20.4% and 20.5% respectively. These exploratory findings are in line with our findings related to people's absolute judgments.

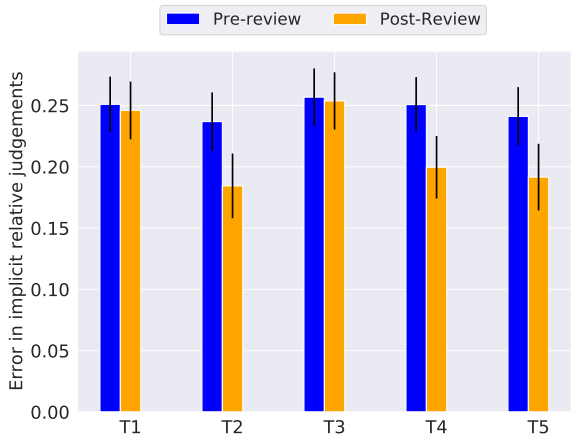
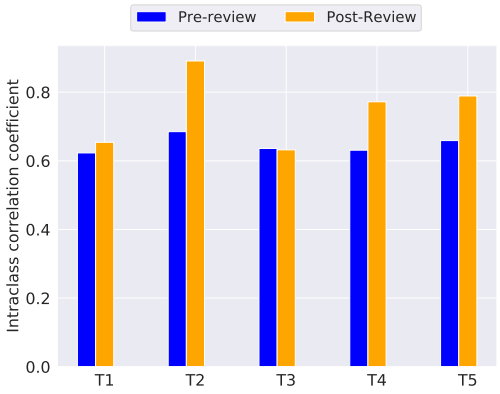


Fig. 8. Error in people's implicit relative judgments. The y-axis shows the fraction of instances where people's implicit relative ordering of apartments (>,< or =) did not match the ground truth ordering based on the listing price. We report mean values calculated across all respondents and pairs of apartments ± 1.96 standard errors of the mean (SEM).

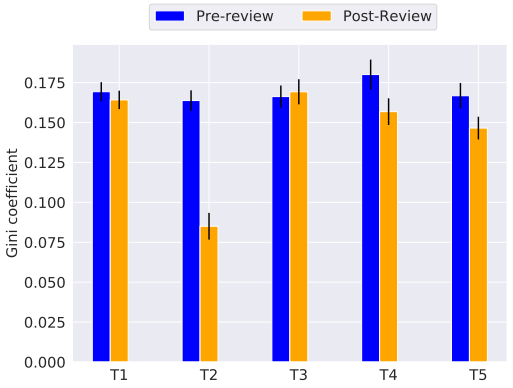
A.2 Other Measures of Inconsistency: an Exploratory Analysis

In Section 4, we quantified consistency as the degree of agreement between respondents' estimates. In this section, we investigate whether our findings hold for a broader set of measures of inter-annotator consistency.

We explore the consistency of two types of dependent variables: the respondents' *absolute* judgments, and their implicit *relative* judgments. In the former, we quantify the consistency between respondents' estimates of apartment prices. In the latter, we focus on the consistency of respondents' judgements regarding the relative ranking or ordering of the apartments. That is, we evaluate whether different respondents assign similar relative positions to the apartments.



(a) Intraclass correlation. The y-axis shows the Intraclass correlation among respondents over all the apartments. A higher value indicates a higher degree of consistency.



(b) Gini coefficient. The y-axis shows the Gini coefficient of people's responses averaged across apartments. A lower value indicates a higher degree of consistency. We report mean values ± 1.96 standard errors of the mean (SEM).

Fig. 9. Effect of the interventions on the consistency of people's absolute judgments. The experimental conditions T1–T5 are shown on the x-axis.

We find that our results about the effects of the studied interventions on inter-annotator consistency are robust across a variety of measures. Descriptively, we observe that treatments T1 and T3 do not impact the degree of consistency between respondents, neither in terms of their absolute judgments nor in terms of their implicit relative judgments. On the other hand, treatments T2, T4 and T5 are found to improve both types of consistency notions. For people's absolute judgments, T2 leads to the greatest increase in consistency. For implicit relative judgments, T4 and T5 increase the overall ranking consistency the most, while all three treatments lead to a similarly large increase in pairwise ranking consistency.

A.2.1 Consistency of People's Absolute Judgments.

Intraclass Correlation Coefficient (ICC). The ICC[7] is typically used as a metric to assess annotators' reliability. An estimate by annotator i for apartment a is modelled as $x_{i,a} = \mu + \alpha_i + \beta_a + \epsilon_{i,a}$, where μ is the unobserved overall mean, α_i models the random effect specific to annotator i, β_a represents the random effect due to the features of apartment a, while $\epsilon_{i,a}$ represents the noise. With this model of annotator estimates, ICC is defined as follows:

$$ICC = \frac{\sigma_\beta}{\sigma_\alpha + \sigma_\beta + \sigma_\epsilon}. \quad (1)$$

Here, σ_β represents the variability in the estimates due to differences in the features of the apartments such as their size or number of rooms. σ_α captures the variability resulting from the differences in the scales used by different respondents, e.g., some respondents may consistently provide higher estimates than others. σ_ϵ accounts for the variability arising due to noise in respondents' evaluations.

If the variability in the estimates is predominantly due to apartments' features the ICC value would be high. Conversely, if there is a high variance in the magnitude of the estimates due to differences in scales (σ_α) or noise (σ_ϵ) ICC would be lower. Essentially, the ICC captures how

[7]We calculated the ICC using the Pingouin library [94]: https://pingouin-stats.org/build/html/generated/pingouin.intraclass_corr.html. We report the ICC3 values, which—in line with our setting—assume a fixed set of k respondents for each instance.

responses cluster for each apartment. A value of zero implies that there are no clusters, and each response is likely to be independent. A value of one implies that all the responses are the same.

Descriptively, we observe that the ICC of people's pre-review estimates is similar across all five treatments, as shown in Figure 9a. In treatments T1 and T3 the ICC of people's post-review estimates remained similar to the ICC of their pre-review estimates. However, in treatments T2, T4 and T5 people's post-review estimates exhibited a higher ICC than their pre-review estimates. Namely, we observe an increase of 0.21, 0.14 and 0.13 in T2, T4 and T5 respectively.

It is important to note that although ICC is a consistent statistic, it has a positive bias, i.e., it overestimates the true value. Additionally, it relies on several assumption such as α, β and ϵ having an expected value of zero and β being uncorrelated with α and ϵ. Below, we consider a metric that does not rely on such modeling assumptions: the Gini coefficient.

Gini Coefficient. The Gini coefficient is a measure of dispersion commonly used to quantify inequality within groups, such as wealth inequality within a nation. Unlike the ICC, the Gini coefficient directly focuses on the differences in respondents' estimates for each apartment, without any modeling assumptions. It is defined as follows:

$$G = \frac{\sum_i^k \sum_j^k |x_i - x_j|}{2 \cdot k \cdot \sum_j x_j}, \tag{2}$$

where x_* denotes the estimate by respondent $*$ and k is the number of respondents. This value captures the dispersion of people's responses for a given apartment. A value of zero indicates that the responses are closely clustered together, while a value of one means the estimates are completely dispersed. To quantify the dispersion across all apartments, we calculate the average Gini coefficient across all 30 apartments.

We report our findings in Figure 9b. Descriptively, we found a similar trend as for the ICC. For treatments T1 and T3, the Gini coefficients of people's pre- and post-review estimates remained similar. On the other hand, in treatments T2, T4 and T5, people's post-review estimates exhibit a lower Gini coefficient, with a decrease of 0.08 in T2, and a decrease of 0.02 in T4 and T5.

A.2.2 Consistency of People's Implicit Relative Judgments.

Pairwise Consistency. We consider a measure of consistency analogous to the measure of accuracy described in Appendix A.1. We again derive people's implicit relative judgments from their absolute estimates. For each respondent and for each pair of apartments (A,B), we check if Apartment A was estimated to be more (>), less (<) or equally as expensive (=) as Apartment B. However, instead of comparing a respondent's implicit relative judgment with the ground truth ordering, we compare it to the other respondents' orderings—namely, to the majority vote of others' implicit relative judgments.

In Figure 10a, we show the average degree of disagreement between individual respondents' relative judgments and the majority vote. We observe that people's pre-review estimates are similarly consistent across all experimental conditions. In treatments T1 and T3, people's post-review estimates exhibit a similar degree of pairwise consistency as their pre-review estimates. However, in treatments T2, T4 and T5 we find that the average disagreement with the majority vote is decreased by 7.5, 6.9 and 7.1 percentage points respectively. It is important to note that the pre-review disagreement was already quite low, leaving little room for improvement. The observed decrease in disagreement in T2, T4 and T5 respectively correspond to 38%, 32% and 35% of the total possible decrease.

While this metric focused on pairwise consistency, below we consider a metric that quantifies the consistency in the overall ordering of the apartments: Kendall's W.

(a) Pairwise consistency. The y-axis shows the average degree of disagreement between individual respondents' relative judgments and the majority vote. A lower value indicates a higher degree of consistency. We report mean values ± 1.96 standard errors of the mean (SEM).

(b) Kendall's W. The y-axis shows the values of Kendall's W statistic calculated on the respondents' implicit rankings. A higher value indicates a higher degree of consistency.

Fig. 10. Effect of the interventions on the consistency of people's implicit relative judgments. The experimental conditions T1–T5 are shown on the x-axis.

Kendall's W. In order to assess the consistency of respondents' overall ordering of apartments, we treat the provided price estimates as implicitly ranking all of the apartments from the least expensive to the most expensive, allowing for ties. We then quantify the consistency between the respondents' implicit rankings utilizing Kendall's W,[8] a non-parametric statistic for rank correlation.

Kendall's W is commonly employed to evaluate agreement amongst respondents in ranking tasks. At a high level, Kedall's W corresponds to the normalized sum of squared deviations from the mean in the rankings. A value of one would indicate perfect agreement amongst respondents, while a value of zero would indicate no agreement.

In Figure 10b we show the values of Kendall's W statistic calculated on the respondents' implicit pre-review and post-review rankings. Descriptively, we observe that in treatments T1 and T3, the respondents' pre-review and post-review rankings are consistent to a similar degree. However, treatments T2, T4 and T5 show an increase in Kendall's W of 0.12, 0.16 and 0.15 in the post-review rankings compared to the pre-review rankings.

[8]We used the Pingouin library [94] to compute Kendall's W: https://pingouin-stats.org/build/html/generated/pingouin.friedman.html. Please note that Kendall's W is computed with a correction for ties.