

INCENTIVES, FRAMING, AND RELIANCE ON ALGORITHMIC ADVICE: AN EXPERIMENTAL STUDY*

BEN GREINER[†], PHILIPP GRÜNWALD[‡], THOMAS LINDNER[§],
GEORG LINTNER[¶], MARTIN WIERNSPERGER^{||}

ABSTRACT

Managerial decision-makers are increasingly supported by advanced data analytics and other AI-based technologies, but are often found to be hesitant to follow the algorithmic advice. We examine how compensation contract design and framing of an AI algorithm influence decision-makers' reliance on algorithmic advice and performance in a price estimation task. Based on a large sample of almost 1,500 participants, we find that compared to a fixed compensation, both compensation contracts based on individual performance and tournament contracts lead to an increase in effort duration and to more reliance on algorithmic advice. We further find that using an AI algorithm that is framed as incorporating also human expertise has positive effects on advice utilization, especially for decision-makers with fixed pay contracts. By showing how widely used control practices such as incentives and task framing influence the interaction of human decision-makers with AI algorithms, our findings have direct implications for managerial practice.

Keywords: artificial intelligence, algorithmic advice, human-augmented algorithmic advice, trust, financial incentives, decision-making

*We thank Christoph Feichter, Razvan Ghita, Wenquian Hu, Steve Kachelmeier, Andrew Schotter, Bei Shi, Gerhard Speckbacher, four anonymous reviewers, as well as seminar, workshop, and conference participants at the ENEAR conference in Seville, the ESA conference in Bologna, the 13th Conference on New Directions in Management Accounting in Lisbon, the 2023 MAS Midyear Meeting in Atlanta, the IMT School for Advanced Studies in Lucca, the University of Hamburg, the University of Innsbruck, UNSW Sydney and WU Vienna for valuable comments and suggestions.

[†]Wirtschaftsuniversität Wien, Institute for Markets and Strategy, Austria, e-mail: bgreiner AT wu.ac.at, and University of New South Wales, School of Economics.

[‡]Wirtschaftsuniversität Wien, Institute for International Business, Austria, e-mail: philipp.gruenwald AT s.wu.ac.at.

[§]University of Innsbruck, Institute for Strategic Management, Marketing and Tourism, Austria, e-mail: thomas.lindner AT uibk.ac.at, Copenhagen Business School, and Wirtschaftsuniversität Wien.

[¶]Wirtschaftsuniversität Wien, Institute for International Business, Austria, e-mail: georg.lintner AT s.wu.ac.at.

^{||}Cornell University, Samuel Curtis Johnson Graduate School of Management, Ithaca, New York, e-mail: martin.wiernsperger AT cornell.edu.

I INTRODUCTION

The emergence of big data has let algorithms and artificial intelligence (AI) enter the everyday activities of businesses and other organizations. While low-stakes routine tasks are increasingly being automated by advanced data analytics and other AI-based technologies, high-stakes managerial decision-making has so far been largely spared from this trend (Keding and Meissner, 2021; Wilson and Daugherty, 2018). Although managers are typically not replaced by algorithms, they are frequently supported by technological decision-aids in order to make better informed and more efficient decisions, which can contribute to building competitive advantage over competitors with less algorithmic support in decision-making (Allen and Choudhury, 2022; Choudhury et al., 2020; Krakowski et al., 2023; Raisch and Krakowski, 2021). However, despite the rapidly improving accuracy of AI algorithms (Dellermann et al., 2019), a growing number of empirical studies observe that human decision-makers are more likely to rely on their own judgement or expert opinions than the advice generated by algorithms, a phenomenon called “algorithm aversion” (Commerford et al., 2022; Dietvorst et al., 2015, 2018).

While previous research has explored how task characteristics (Castelo et al., 2019; Hertz and Wiese, 2019), the ability to modify an algorithm (Costello et al., 2020; Kawaguchi, 2020), or various personality traits (e.g., Cao et al., 2022; Dietvorst and Bharti, 2020) influence algorithm aversion, we know little about how decision-makers’ incentives and the possible incorporation of human-expert knowledge into AI advice influence how decision-makers deal with algorithmic advice. In this paper, we attempt to contribute to filling this gap by experimentally examining the causal effect of the design of decision-makers’ compensation contracts and the framing of an AI algorithm on advice utilization, the decision-makers’ effort duration, and eventual performance.

From a traditional economics perspective, self-interested decision-makers, when being exposed to performance-based incentives as opposed to rewards not related to performance, should rely more on a human-outperforming algorithmic decision aid instead of relying on own (effortful) judgements. However, a three-decades old but still prominently cited stream of experimental research suggests that incentivizing decision-makers in the presence of an algorithmic decision aid *negatively* influences their performance (Arkes et al., 1986; Ashton, 1990). Accordingly, the provision of financial incentives can “backfire” by encouraging decision-makers to exert unproductive effort instead of relying on the algorithmic advice (Camerer and Hogarth, 1999).

However, several developments suggest a reexamination of these previous results. First, due to the more frequent employment of AI-based decision aids in firms and everyday life over the past decades, attitudes towards algorithms may have changed. Novel methodologies

of aggregating unstructured data emerged since the publication of this older empirical literature. Second, the use of incentives in compensation contracts of managers, auditors, financial analysts, and other upper-level organizational decision-makers is becoming more widespread nowadays, potentially changing their effects on algorithm uptake. Third, most previous studies in this context have tied decision-makers’ compensation not to absolute but rather to relative (tournament) performance (Arkes et al., 1986; Ashton, 1990; Samuels and Whitecotton, 2011). Under tournament contracts, decision-makers are not necessarily incentivized to make good judgements, but rather to outperform fellow competitors, which may be partially driving the observed performance effects. Artificial intelligence algorithms also do not exist in a vacuum, as they are developed and trained by human experts and thus typically incorporate human expertise input when providing advice. If and how decision-makers react differently when knowing that human experts were involved in the generation of algorithmic advice, and whether financial incentives might have a different effect under those circumstances, remains an open question.

Thus, our study intends to shed further light on how different incentive contracts influence decision-makers’ interactions with modern AI algorithms. More specifically, we empirically test the behavioral arguments about the “backfiring” of financial incentives as brought forward by Arkes et al. (1986) and Ashton (1990) against more mainstream economic arguments of rational decision-making. In contrast to other more recent experimental studies on incentives and algorithm use (e.g., Neumann et al., 2022; Samuels and Whitecotton, 2011), we directly compare how different compensation contract designs (fixed payment, performance-based incentives, and tournament incentives) influence decision-makers’ advice utilization from differently framed AI algorithms. In addition, we observe the time decision-makers invest to analyze contextual information and their eventual task performance. This approach allows us to empirically examine if and when the provision of financial incentives can exacerbate or mitigate algorithm aversion.

In our experiment, about 1,500 participants estimate the price per night of multiple Airbnb apartments in Vienna. Participants receive the actual apartment listing information, which includes the cover photo, textual descriptions, a city map with the approximate location, and customer review scores. In addition, they are provided with imperfect price predictions by an AI algorithm, which, with a 30% average deviation, outperforms typical human price estimations. The algorithmic predictions are based on a random forest model and the knowledge of five local renting market experts. In a 3×3 factorial between-subjects design, we manipulate the type of compensation contract and the framing of algorithmic advice. In particular, participants are compensated with either a fixed payment, incentives based on individual performance, or tournament incentives. In terms of algorithmic advice, participants either receive no algorithmic price predictions, price predictions with a description focusing on the random forest model of the AI algorithm, or price predictions with a description highlighting the involvement of human

experts. As main dependent variables, we measure participants’ use of the algorithm (weight of advice), their exerted effort duration, and their task performance (deviation from actual apartment prices).

In our large sample, we find that both types of incentive contracts lead to a substantially higher reliance on algorithmic advice compared to a fixed payment (although we observe a marginally higher advice utilization with individual performance contracts compared to tournament contracts). Further, we observe that both performance-based and tournament incentives have similar positive effects on effort duration, and positive but statistically not significant effects on performance, relative to a fixed payment. Thus, we observe no evidence that incentives undermine decision-makers’ use of algorithmic advice. Concerning our human-framed AI advice treatment, we find that participants with a fixed payment rely more on algorithmic advice when knowing that also human experts were involved in the predictions of the AI algorithm. However, such human expert framing of the AI algorithm does not give an additional boost to the advice utilization of decision-makers with performance-based or tournament incentives.

We thereby contribute to different academic conversations on algorithm use in contemporary research in various fields such as accounting, economics, management, and psychology. First, our research highlights under which circumstances human decision-makers and AI algorithms can collaborate to augment each other’s strengths, and is thus related to an emerging literature on human-machine interactions in augmented decision-making, both in accounting (e.g., Costello et al., 2020; Emmett et al., 2023; Estep et al., 2023; Labro et al., 2023; Libby and Witz, 2024; Liu, 2022), and the broader management literature (e.g., Allen and Choudhury, 2022; Choudhury et al., 2020; Raisch and Krakowski, 2021).

Second, numerous contemporary studies found that decision-makers frequently distrust algorithms, in particular when the algorithm makes mistakes from time to time (e.g., Chen et al., 2022; Dietvorst et al., 2015; Prahla and Van Swol, 2017). For instance, in an auditing context, Commerford et al. (2022) and Cao et al. (2022) show that decision-makers trust algorithmic advice less than human expert advice. In contrast to previous research, we do not manipulate human versus algorithmic advice but rather show that simply mentioning the involvement of human experts in the development of an AI algorithm can increase decision-makers’ reliance on advice, especially when they have fixed pay contracts. Thereby, we add a new perspective to the literature on human trust in AI (e.g., Glikson and Woolley, 2020).

Third, our study answers recent calls for more research into the effectiveness of financial incentives in decision-making tasks with algorithmic advice (e.g., Burton et al., 2020; Neumann et al., 2022; Zellner et al., 2021). Older, but still prominently cited research in accounting and psychology highlights the “paradoxical” phenomenon that incentives undermine potentially positive effects of algorithmic decision-aids (Arkes et al., 1986; Ashton, 1990). In line

with other more recent experimental studies (Neumann et al., 2022), we do not observe such incentive-induced algorithm aversion in our contemporaneous setting with a contextually-rich price estimation task, a modern random forest algorithm, and salient financial incentives. Based on our study, we rather conclude that financial incentives increase both effort duration and use of algorithmic advice. Thus, we contribute to a long-standing stream of accounting research on financial incentives in judgment and decision-making tasks (e.g., Awasthi and Pratt, 1990; Ding and Beaulieu, 2011; Farrell et al., 2014; Libby and Lipe, 1992; Libby and Luft, 1993).

Our study also provides guidance for managerial practice. It has been received wisdom that using financial incentives to encourage the use of algorithmic advice may backfire (Arkes et al., 1986; Ashton, 1990). Given vast and ongoing changes to the kinds of algorithmic support systems available to managers, our study re-investigates in a modern design how much risk of backfiring of algorithmic advice still remains today. We find that there is little evidence of incentives working against the uptake of algorithmic advice, for both main incentive types, individual and tournament incentives. We follow up on this insight and compare algorithmic advice to advice that joins human and algorithmic insights. Such advice is more readily taken up by participants in our experiments.

As such, our study offers two actionable strategies to mitigate algorithm aversion. AI tools are increasingly used at all organizational levels, from CEOs and other senior managers to shop-floor employees. While the use of incentives in compensation contracts for higher-level managerial decision-makers is almost ubiquitous, regular employees making day-to-day operational decisions with the support of AI tools typically receive a fixed salary. Our study highlights that these decision-makers are more prone to algorithm aversion and that organizations can benefit from offering them financial incentives. Our second recommendation is that when it is not possible or practically feasible to use financial incentives, a simple framing intervention highlighting the involvement of human experts in the development of the AI decision tool can also positively affect attitudes towards algorithmic advice. We thereby add a dimension to managerial insights into hybrid intelligence which so far emphasized the interpretability of algorithms and the anthropomorphic advice provided by language models (e.g., Kellogg et al., 2020; Murray et al., 2021; Shrestha et al., 2019).

Our paper proceeds as follows. In Section II we review previous research on the effect of incentives on algorithm use, and derive hypotheses for our experimental study. Section III describes our experimental design and procedures. Section IV presents the data analysis and discusses our results, while Section V concludes.

II LITERATURE REVIEW AND HYPOTHESIS DEVELOPMENT

Previous literature has established that decision-support from algorithmic advice can, in many circumstances, enrich human decision-making (Dellermann et al., 2019; Estep et al., 2023). Due to recent improvements, algorithmic advice has become superior to human expert advice in many decision-making settings (e.g., Choudhury et al., 2020; Dietvorst and Bharti, 2020; Labro et al., 2023), and decision-makers should disregard algorithmic advice only when their human intuition, tacit knowledge, and experience allows them to interpret contextual circumstances in a superior way (Raisch and Krakowski, 2021). The field of medicine has been a particularly fertile ground providing evidence for superior diagnostic performance combining human cognition and algorithmic advice (e.g., Goldstein et al., 2017; Rajpurkar et al., 2022; Tschandl et al., 2020). Recent insights from management and related fields include benefits of algorithmic decision-support for strategy development (e.g., Dell’Acqua et al., 2023; Krakowski et al., 2023), performance monitoring (e.g., Labro et al., 2023), and specialist training (e.g., Gaessler and Piezunka, 2023).

However, the relationship between algorithmic advice and decision quality seems to rest on several contingencies. Humans tend to overestimate their own decision-making capabilities, which leads to inferior performance (see e.g., Hoffman et al., 2017, in a hiring context). A lack of trust in decision-support systems (Wang et al., *ming*) will reduce the benefits of algorithmic advice, and decision-makers are less likely to rely on algorithmic advice when the advice does not align with past experience (Liu et al., *ming*). When the algorithms underlying decision-support systems are intransparent, decision-makers are also less likely to employ algorithmic advice (e.g., Bauer et al., 2021, 2023; Poursabzi-Sangdeh et al., 2021). The distrust in algorithmic advice seems to be task-dependent. For instance, Castelo et al. (2019) show that algorithmic decision aids are trusted less for tasks that seem rather subjective versus objective in nature. Hertz and Wiese (2019) find that people trust advice from an algorithmic source more when working on analytical tasks than social tasks. In a hiring context, Dargnies et al. (*ming*) show that providing overconfident managers with feedback on their past hiring performance increases their voluntary adoption of algorithmic advice, while providing more details about how the algorithm works does not. Finally, in an experiment, Jung and Seiter (2021) find that algorithm aversion vanishes when decision-makers are working under time pressure.

II.A Algorithmic Advice and Financial Incentives

Despite their relevance for the proliferation of new technologies in general, research on the relationship between financial incentives and the use of algorithmic advice in decision-making still relies on a set of rather traditional studies. Although the effortful task of combining

algorithmic advice with human judgment would intuitively benefit from incentivizing human decision-makers (Burton et al., 2020), a “backfiring” of financial incentives has been observed in the presence of algorithmic advice. Experimental studies by Arkes et al. (1986) and Ashton (1990) find that, in the presence of algorithmic decision aids, incentivized participants perform worse on judgment and decision-making tasks than unincentivized participants.

To explain this “paradoxical” effect, Ashton (1990) invokes a behavioral argument. Financial incentives increase performance pressure on a decision-maker, and the effect of this pressure depends on the nature of the decision task, e.g., whether it is boring and monotonous, or interesting and requires complex cognitive activities.¹ Ashton (1990)’s argument is that the introduction of an algorithmic decision aid can change the nature of a task, in that an otherwise dull task becomes interesting and challenging, inducing a (misguided) belief that it requires customized solutions to achieve high performance, as opposed to simply following the algorithmic advice. Put differently, incentivized decision-makers feel as if they have to earn their reward by coming up with their own judgments instead of using the readily available algorithmic advice. Similarly, Camerer and Hogarth (1999) as well as Awasthi and Pratt (1990) suggest that financial incentives may motivate decision-makers to exert too much, misdirected effort. Samuels and Whitecotton (2011) suggest that both the size and the direction of the “backfiring” effect of financial incentives may depend on the amount of contextual information available to decision-makers in addition to the algorithmic advice.

This finding, prominently discussed in the managerial accounting literature, contrasts economic theory as well as other experimental evidence on the effect of economic incentives on performance. According to basic economic theory, self-interested individuals would always work harder and more effectively if their compensation is properly tied to performance. The use of performance incentives for managers, auditors, financial analysts, and other upper-level organizational decision-makers is common. In their review of experimental studies on judgment and decision-making tasks, Camerer and Hogarth (1999) report that the provision of financial incentives generally has positive effects on performance by improving the recall of remembered items, mitigating anchoring bias, and reducing the variance of decision quality. The same has been found in managerial contexts. Sprinkle (2000) reports performance-increasing effects of incentives in task with production output decisions, in particular when subjects have the ability to learn over multiple periods. Ding and Beaulieu (2011) show that in a balanced-scorecard based judgment task, the provision of financial incentives reduces the unintended influence of decision-makers’ affective biases (e.g., mood, emotions) on decision outcomes. Similarly, in a series of experiments, Farrell et al. (2014) collect behavioral and brain-activity data showing that

¹Accordingly, Bonner et al. (2000) find that the likelihood of observing positive effects of financial incentives on performance decreases in the complexity of the task and required skills.

performance-based incentives induce decision-makers to process information more analytically and to make more economically desirable investment choices.

These two conflicting streams of theoretical explanations and evidence, in combination with rapid developments in the nature and application of algorithms in recent years, motivate our study design. Using treatment conditions with and without algorithmic advice, and with and without economic incentives, we test the following competing hypotheses. Hypothesis 1a follows previous experimental “backfiring” evidence and associated behavioral arguments (Arkes et al., 1986; Ashton, 1990; Camerer and Hogarth, 1999) where the provision of financial incentives causes decision-makers to exert more unproductive effort and to rely less on algorithmic advice, which eventually undermines their performance. This perspective is contrasted in Hypothesis 1b with the economic theory argument that financial incentives should lead to more algorithm use and better performance, given a reasonably powerful algorithm that outperforms an average human decision-maker.

Hypothesis 1a. *Decision-makers with financial incentives rely less on algorithmic advice (and perform worse) than decision-makers who receive a fixed payment.*

Hypothesis 1b. *Decision-makers with financial incentives rely more on algorithmic advice (and perform better) than decision-makers who receive a fixed payment.*

II.B Types of Financial Incentives

Conditional on incentives being provided, the degree to which incentives contribute to the uptake of algorithmic advice (and eventually to better performance) likely depends on the type of financial incentives. To date, most experiments on judgment and decision-making with algorithmic advice have tied compensation not to absolute but rather to relative performance (e.g., Arkes et al., 1986; Ashton, 1990; Samuels and Whitecotton, 2011). With tournament incentives, decision-makers are not necessarily incentivized to make good judgements but rather to outperform fellow competitors (Burton et al., 2020; Lazear and Rosen, 1981). Already Ashton (1990) argues that in a tournament, knowing that their peers have access to the same algorithmic advice, decision-makers may assume that merely relying on the algorithmic advice will not be sufficient to secure a top position. This perception could encourage decision-makers to develop their own solutions or to use heuristics, rather than following the provided algorithmic advice.

Ottaviani and Sørensen (2006) formalize this intuitive argument in their “forecasters” model, where each forecaster receives a private signal about the state of the world (e.g., based on examination of available information), but also has access to a common signal (in our context: the algorithmic advice). They show that, compared to individual performance incentives, in

a tournament setting expected-utility-maximizing agents will put more weight on their own signal, as they now aim to maximize the likelihood to win against the other agent rather than to maximize their forecast accuracy. Intuitively, if both agents share the same public signal, the only way to set oneself apart is by exploiting a private signal. This incentive is present not only for risk-loving individuals, but also for risk-neutral and slightly risk-averse ones.

These arguments are the basis for our design choice to explore two types of incentives: individual performance-based payments, and a tournament. Hypothesis 2 suggests that under tournament incentives, decision-makers are less likely to rely on algorithmic advice. As a consequence, if algorithmic advice is superior in quality to the (average) human decision-maker, absolute performance will decrease.

Hypothesis 2. *Decision-makers with tournament incentives rely less on algorithmic advice (and perform worse on average) than decision-makers with performance-based incentives.*

II.C Types of Algorithmic Advice

A stream of (experimental) research has compared decision-makers’ reactions to human vs. algorithmic advice.² Dietvorst et al. (2015) report that decision-makers lose confidence in an algorithmic decision aid more quickly than in a human expert upon observing its mistakes. Similarly, Efendić et al. (2020) show that decision-makers judge slowly generated advice from algorithmic decision aids to be of lower quality than slowly generated human advice. Dietvorst and Bharti (2020) find that decision-makers favor riskier, and often worse-performing, human judgment whenever they feel that an algorithm is unlikely to give near-perfect advice. In an applied accounting setting, Commerford et al. (2022) show that auditors, who receive contradictory evidence from an AI decision aid (instead of a human specialist), rely less on the advice when proposing audit adjustments. Similarly, Chen et al. (2022) observe that managers perceive negative sales forecasts as being less credible when they come from algorithms than human experts.³

In modern algorithms, however, the distinction between algorithmic and human advice becomes blurry. Algorithmic advice is often processed and modified, by the agent herself or by other people (e.g., team members or assistants), before it enters the decision-making deliberations. For example, a laboratory experiment by Dietvorst et al. (2018) as well as a field experiment by Kawaguchi (2020) observe that giving human decision-makers the possibility to

²For an overview of how people deal with *human* advice in various conditions and under various incentives, see Schotter (2023).

³One exception is Logg et al. (2019), who find that people adhere more to advice when they think it comes from an algorithm than from a human individual. This “algorithm appreciation” vanishes, though, when people have to choose between their own advice and algorithms, or when they have more knowledge about the task.

modify an algorithmic advice substantially increases their propensity to use the algorithmic decision aid, which can have a positive effect on their performance.

For these reasons, in our study we do not only employ a purely algorithmic advice tool, but also “humanized” algorithmic advice that is prominently framed as including human expert input. We thus test whether, holding algorithmic performance and advice constant, an algorithmic tool that is augmented by human expertise is trusted more than an AI algorithm that is framed as only based on machine input. Adding this additional type of algorithmic advice allows us to test the robustness of effects of financial incentives on algorithm uptake (be they positive or negative) across differently framed algorithms. It also allows us to explore whether higher trust in a humanized algorithm holds under different incentive conditions, since previous literature comparing algorithmic and human advice has often used fixed payments or no payments at all (e.g., Castelo et al., 2019; Chen et al., 2022; Commerford et al., 2022).

Our Hypothesis 3 extrapolates previous findings of higher trust in human than algorithmic advice to our humanized algorithmic advice, thus postulating that framing the advice as based on both human and AI input reduces algorithm aversion, leading to higher reliance on the algorithmic advice and, consequently, higher performance.

Hypothesis 3. *Decision-makers rely more on algorithmic advice (and perform better) when the algorithm also considers human expertise.*

Our full-factorial design of an incentive dimension (no incentives, individual performance incentives, tournament incentives) and an advice type dimension (no advice, algorithmic advice, humanized algorithmic advice) allows us to explore the robustness of the effects of providing monetary incentives across different advice types, and of changing the framing of algorithmic advice across different incentive conditions. However, it is difficult to predict possible interaction effects ex-ante. First, there is a lack of previous research upon which we could build our prediction. Second, we have a set of competing hypotheses for the effect of financial incentives, and moderating effects of humanized algorithmic advice may differ depending on direction and mechanism of incentive main effects. For instance, if financial incentives induce decision-makers to maximize earnings and thus to focus on the prediction error of the algorithm, we may expect that pure framing in the description of the advice will have less of an effect. As a result, the human-expert input framing could have a stronger effect on decision-makers with a fixed payment than on decision-makers with performance-based or tournament incentives. If, on the other hand, financial incentives induce decision-makers to exert unproductive effort and to generally distrust algorithmic advice, the human-expert framing of the algorithm could counteract this negative effect. Thus, we will treat possible interactions as an exploratory analysis, charting new territory without a specific hypothesis.

III EXPERIMENTAL DESIGN AND PROCEDURES

Our experiment employs a 3×3 factorial between-subjects design in order to study the effect of individual and tournament incentives as well as human-expert-framing of algorithmic advice on decision-makers’ algorithm use, their exerted effort duration, and their task performance.

III.A Task

Participants had to estimate the price per night of Airbnb apartments in Vienna (Austria) as of June 2021. Cost-, price-, demand-, and revenue-forecasting are part of many managerial occupations, and are activities that will benefit greatly from the advancement of algorithms.⁴ Out of a sample of 11,567 listings obtained from the open-source project *Inside Airbnb* (www.insideairbnb.com), 10 apartment listings were selected which were no longer publicly available on the Airbnb platform at the time of the experiment (so prices could not be looked up through a search). For each listing, participants were provided with a substantial amount of contextual information from the original listing.⁵ Each participant received the 10 listings in random order, one by one. In order to reduce spillovers both between tasks and between participants, our experimental subjects did not receive feedback about their performance, neither in-between tasks nor at the end of the study.

III.B Independent Variable 1: Compensation Contract Design

To examine how the design of decision-makers’ compensation contracts influences their use of the algorithmic advice and their eventual task performance, we manipulate how participants get paid. Participants, if selected to be paid (see below for details on our implementation of a between-subjects random incentive system), either received a performance-independent fixed payment, individual performance-based incentives, or tournament incentives.

More specifically, in the *fixed payment* condition, participants received a lump sum of EUR 50 for completing the price estimation task, not contingent on their task performance. Participants in the *performance-based incentives* condition were paid according to an incentive-compatible binarized quadratic scoring rule (Hossain and Okui, 2013). They received either a payment of EUR 100 or EUR 0, with the probability of the large prize

⁴Poursabzi-Sangdeh et al. (2021) use a similar price forecasting task to explore the effect of model complexity, and Chen et al. (2022) employ a demand forecasting task in their study.

⁵This included original listing title, cover photo, room type (entire apartment or private room), district of Vienna and an approximate location on a city map, number of accommodated guests, number of bedrooms, number of beds, number of bathrooms, superhost status, identity verification status, number of reviews, average overall review rating, average review rating within six categories (Accuracy, Cleanliness, Check-in, Communication, Location, and Value), and the original “About this space” description (limited to 500 characters). See Appendix E for screenshots of the task presentation.

being (quadratically) contingent on their performance in the price estimation task, equaling $\max\{100 - 0.2 \times (\text{estimate} - \text{true price})^2, 0\}$. Theoretically, this payment rule also neutralizes risk attitudes (of expected utility maximizers).⁶

In the third, *tournament incentives* condition, participants were randomly paired with a second participant to determine their relative performance. In each pair, the better performing participant (i.e., the one with the lower deviation from the true apartment price) received a payment of EUR 100, while the worse performing participant received EUR 0.

III.C Independent Variable 2: Algorithmic Advice and its Framing

To examine if and how the effect of financial incentives might be different in the absence versus the presence of an algorithmic decision aid, we manipulate between-subjects whether participants receive advice from an AI algorithm or not, and how the inclusion of human expertise in the algorithmic advice is framed.

In the *no advice* condition, participants worked on the price estimation task only based on the provided contextual information, without any algorithmic decision aid. In the conditions *AI advice* and *human-AI advice*, participants had to submit two estimates for each of the 10 listings: first without any advice using only the contextual information (as in the *no advice* condition), and then again after receiving a prediction from an algorithm.⁷

To provide participants with algorithmic advice, we developed a random forest model that utilizes a raw dataset of 11,567 apartment listings in Vienna and generates price predictions based on numerous numerical input variables.⁸ In addition, we obtained price estimates for the selected apartments from five experts. These experts were active landlords in Vienna with substantial experience with the Viennese real estate sector and in professionally renting out

⁶For implementation, we randomly drew a number between 0 and 100. If the score was higher than or equal to that random number, the participant received EUR 100, and EUR 0 otherwise. In theory, this approach allows us to elicit subjects' truthful beliefs independent of their risk attitudes. See Schotter and Trevino (2014), Schlag et al. (2015), or Charness et al. (2021) for reviews of theoretical and empirical evidence on incentive-compatible scoring rules. To ease understanding of the binarized scoring rule, we additionally provided a table mapping different estimation errors to probabilities for the high prize, and explained that the more accurate a participant's price estimations, the higher is the chance to receive the EUR 100 reward.

⁷The two-stage design was mainly a methodological choice to cleanly measure the weight of advice in the final estimate. However, we believe that the procedure also reflects managerial practice well. When facing a decision, financial analysts, auditors, and managers in various other functional domains typically start with receiving or collecting relevant contextual information, forming their own initial (sometimes intuitive) judgment. This initial step is then often followed by the consultation of algorithmic decision aids. The use of AI-based tools usually requires some prior knowledge of the subject matter. At the same time, in managerial decision-making and many other settings, it is very unlikely that the advice from an AI system is the final decision. Recent legislation proposals in the European Union even include a mandate for human intervention whenever there are legal ramifications of a decision (e.g., in HR and hiring contexts).

⁸Specifically, these are apartment type, number of bedrooms, number of beds, number of accommodated guests, district of Vienna, number of reviews, average review rating, and superhost status. The model and its selection process are described in detail in Appendix D.

apartments in Vienna through the Airbnb platform. They completed the task in advance, without an algorithmic decision aid. The final algorithmic decision aid displayed a weighted average of those two estimates, with the random forest model weighing 80% and the average human expert estimate weighing 20%. Participants knew that this algorithmic aid had a mean absolute error of about 30% and that it derived its price predictions based on several components. In our context, the algorithm easily outperforms an average human decision-maker.⁹

The two algorithmic advice conditions only differ in the *framing* of the algorithm. In the *human-AI advice* condition, we de-emphasized the random forest model, but highlighted the human expert involvement.¹⁰ However, note that in both conditions, participants received exactly the *same* algorithmic advice and the same information that it has a mean absolute error of 30%. For a rational Bayesian decision-maker, only the eventual precision of the given advice is relevant, not on how many and which sources it relies. In this sense, our *human-AI advice* condition is purely a framing manipulation.^{11,12}

III.D Dependent Variables: Algorithm Use, Effort, and Estimation Error

Our three main dependent variables are participants’ use of the algorithm (measured as the *weight of advice*), their exerted effort duration (measured as *time* in seconds), and their performance (measured as the absolute *estimation error*).

First, for those participants who receive algorithmic advice, we measure how much they rely on this advice when making price estimations. In line with previous experimental research on advice-taking in judgment tasks (e.g., Logg et al., 2019; Prahla and Van Swol, 2017; see

⁹The relative mean absolute error of participants in our no-advice condition is 76.50%, with 95% of participants performing worse than the algorithm.

¹⁰In particular, in the human-AI advice conditions we describe to participants that “the price estimate incorporates the expert advice from 5 individuals. The five experts have substantial experience in the pricing of Airbnb apartments and are familiar with the housing and accommodation sector in Vienna.” We intentionally kept the expert description rather vague and only refer to their expertise, in order to mitigate potential effects of prior positive or negative experiences with landlords.

¹¹The framing may also affect the perception of parameters of the advice error distribution which we did not fix through the instructions. For example, participants may think that a combination of algorithmic advice with human expertise may curtail long tails of the error distribution. However, for typical symmetric single-peaked mean-zero error distributions, the effect of presumed affected secondary distribution parameters (kurtosis, skewness) while holding the mean absolute error constant on participants’ guesses is likely negligible.

¹²Following the recommendation of a reviewer, we ran a follow-up study in which we tested how decision-makers judge the credibility of advice from five additional types of human-AI algorithm interactions. In this additional experiment with more than 1,500 participants, we do not find any meaningful differences in advice credibility between the relatively vague description of a human-AI algorithm used in our experiment, and five additional, more detailed descriptions of possible human-AI algorithm interactions. We present this additional evidence in Appendix C.

Bonaccio and Dalal, 2006, for an overview), we measure participants’ algorithm use as the *weight of advice* based on a definition established by Harvey and Fischer (1997), by relating the absolute difference between participants’ final and initial price estimation to the absolute difference between the algorithmic advice and participants’ initial price estimation. Since, as Bonaccio and Dalal (2006) discuss, this definition may lead to ambiguous values smaller than 0 or larger than 1 in certain cases (e.g., when initial and final estimate are on different sides of the algorithmic advice, or the decision-makers’ final price estimation moves in the opposite direction of the advice), we censor the weight of advice to values between 0 and 1.¹³

$$\text{Weight of advice} = \min \left(\max \left(0, \frac{\text{abs}(\text{final price estimation} - \text{initial price estimation})}{\text{abs}(\text{algorithmic advice} - \text{initial price estimation})} \right), 1 \right)$$

As a second dependent variable, we measure how much effort participants exert on the price estimation task, by collecting data on the *time* (in seconds) that participants devote to estimating the price per night for each Airbnb apartment (“effort duration”, see Bonner and Sprinkle, 2002). Importantly, we kept the length of the information presented on the screens exactly the same across treatments, so that our time measure accurately captures effort duration. For participants receiving algorithmic advice, effort duration is the sum of the time spent on the initial price estimation and the time spent on the final price estimation. For participants in the no advice condition, effort duration measures the time spent on their initial (and only) price estimation. In our analyses, we discuss differences in the time spent in total as well as on the initial and final price estimation.

Third, in line with antecedent research on algorithm aversion (e.g., Dietvorst et al., 2015, 2018), we define participants’ performance in the decision-making task in terms of their *estimation error*, which we measure as the absolute deviation between their final price estimation and the actual listing price of the apartment on the Airbnb platform.

III.E Experimental Procedures

We recruited participants from experimental laboratory subject pools at three large public universities in Austria via the recruitment system ORSEE (Greiner, 2015). Each invited person received a unique invitation link, allowing us track (potential) double participation. We received 1,634 full responses. 117 participants failed an attention check (see below). We excluded a further 28 participants from the analysis due to potential double participation (using the same invitation link, though they may indeed be different subjects) or missing contact details,

¹³We note that when the algorithmic advice is identical to a decision-maker’s initial estimate, the weight of advice is undefined. In total, we have 179 cases of undefined weight of advice, which represents less than 2% of all advice observations. Since there is no established approach to correct those undefined values, we omit them in the respective analyses.

leaving a remaining sample of 1,489 participants for analysis. Table 1 shows the distribution of participants over the nine treatment cells.¹⁴

The mean age of our participants was 25 years, and 61% were female. About 50% are undergraduate students, 40% are graduate students, and the remaining 10% either completed their studies or pursue more advanced postgraduate studies. About 53% of the final sample are Austrian nationals, 16% are German, with the remaining participants being from 64 different countries.

TABLE 1: RANDOM SAMPLE SIZES PER TREATMENT CONDITION

	No Advice	AI Advice	HumanAI Advice
Fixed Pay	N = 169	N = 167	N = 186
Performance Pay	N = 172	N = 170	N = 156
Tournament Pay	N = 159	N = 168	N = 142

We believe that our sample of student participants is suitable for our research question and that it has multiple advantages over samples of professionals. First, our sample fits well to the task of estimating the price of Airbnb apartments, as more than 75% of our participants had booked an apartment via this platform in the past. While participants were familiar with the setting, the task was still sufficiently challenging for them. Second, student participants have low opportunity costs for participating, they can be more cost-effectively incentivized in the experiment, and they have steep learning curves, quickly adapting to the experimental environment. Third, with participants recruited from established offline university subject pools, we can rule out that bots participated in our experiment. Frechette (2015) finds in a review that conclusions reached by using standard student pools mostly generalize to professionals.

Our experiment was self-paced and programmed in Qualtrics. Participants were randomly assigned to one of the 9 experimental conditions. After an informed consent screen, participants received instructions conditional on the assigned treatment (see Appendix E for instructions and screenshots). Before participants proceeded to the actual price estimation task, they had to correctly answer some comprehension check questions.

¹⁴Table A.1 in the Appendix provides a complete balance table across participant characteristics. In our initial power calculation we considered a small-to-medium standardized effect size (Cohen’s d) of 0.4. Arkes et al. (1986) and Ashton (1990) do not provide sufficient statistics in their papers to calculate relative effect sizes, with the exception of Ashton (1990)’s results on the effect of incentives on task performance, which yields a Cohen’s d of 0.574. For a t -test between two treatment conditions at 80% power and with an alpha level of 0.05, and effect size of 0.4 implies a minimum sample size of 99 participants, such that we aimed for 100 participants per cell. In the end we received 165 valid responses per treatment cell on average. This implies an (average) power of 95% to detect the above standardized effect size, or respectively allows us to detect even smaller effects at 80% power.

As another check of proper participation and attention, we added an “attention check listing” to the 10 actual apartment listings, for which all participants needed to submit “1,000” as price per night. Participants were informed that there may be an attention check and that failing to pass it would lead to their exclusion from payoffs.

After completing the main experimental task, participants answered a post-experiment-questionnaire that included measures for demographics, risk-taking attitudes (item based on Dohmen et al., 2011; Tasoff and Zhang, 2022), task enjoyment (intrinsic motivation), overconfidence, information reliance, unfamiliarity challenges, and source credibility (item based on Chen et al., 2022). On the final screen, we also asked them to provide us with their full name and email address as we needed this information to administrate payments.¹⁵

The study ran for about a week in early December 2021. Across all experimental conditions, the median time to complete the experiment was 19 minutes. After the study concluded, we randomly selected 102 participants for payment. These participants earned either EUR 0, EUR 50, or EUR 100, depending on their experimental condition (see above) and their performance in one of their 10 or 20 estimation tasks.¹⁶ On average, these selected participants received a payment of EUR 45.10 via bank transfer. This implies an (ex-post) average payment of EUR 3.09 across the full sample for a 19-minutes task.

IV RESULTS

In the presentation of our results, we start by discussing the distribution of participants’ estimates with respect to the advice given by the algorithm. We then present the outcomes in terms of our three main dependent variables, weight of advice, time spent on the task, and estimation error. We use regressions to analyze differences between our treatment conditions. This is followed by a discussion of responses to our post-experimental questionnaire and their relation to our variables of interest. We conclude this section with a discussion of our results with respect to the hypotheses laid out in Section II.

IV.A *Distribution of Estimates*

Our participants often deviate from algorithmic advice. The normalized histograms in Figure 1 present the relative deviation of participants’ own price estimations from the algorithmic pre-

¹⁵Only one participant did not provide these contact details (and is excluded from the final sample).

¹⁶We thus use a “between-subjects random incentive system”. Theoretical and empirical evidence in experimental economics (see, e.g., Azrieli et al., 2018; Charness et al., 2016) suggests that this is incentive-compatible and preempts wealth effects and hedging strategies, and that participants appear to react more strongly to the nominal value of a payment than to the probability of receiving that payment (see, e.g., March et al., 2016). We pay 102 participants instead of 100 as announced because we needed to form matched pairs in the tournament incentive conditions after random selection of participants.

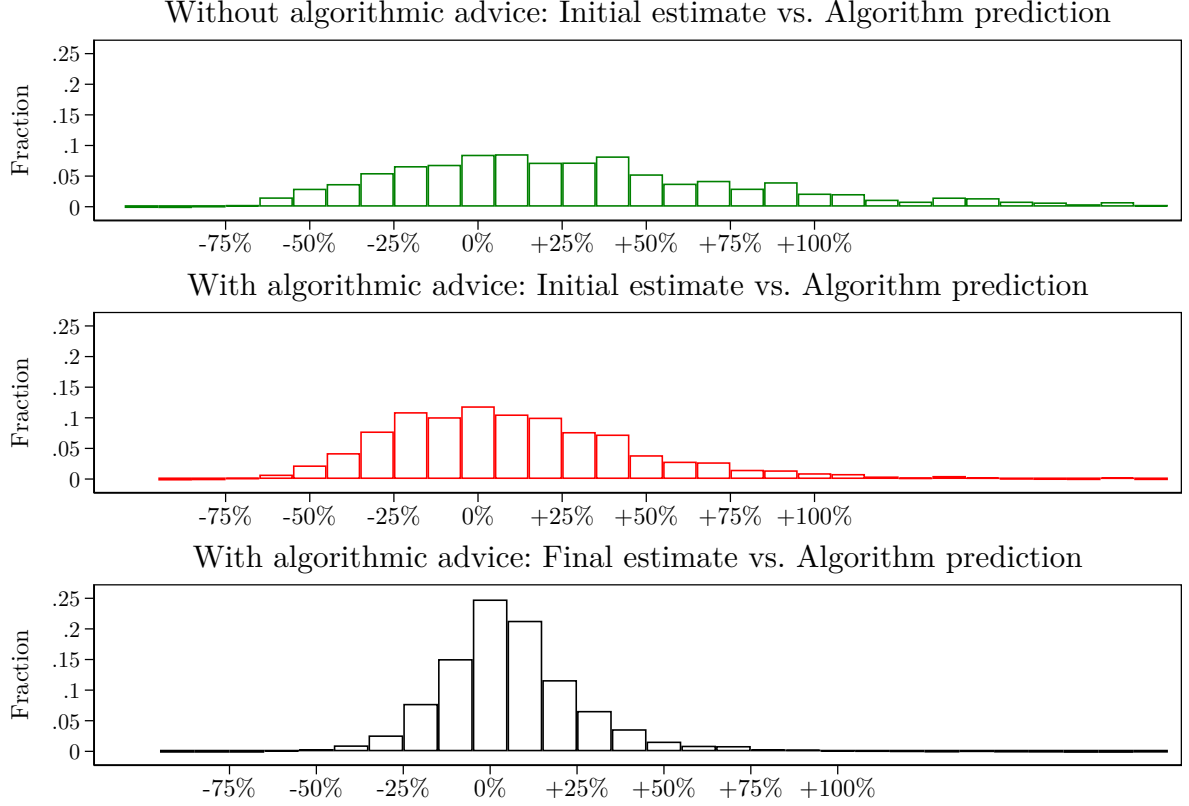
dictions. The first histogram focuses on the estimates of participants who do not receive any algorithmic advice, and thus have to make only one price estimation for each of the 10 listings. We observe that estimates vary considerably. Most of those estimates are in the range of -50% to +100% relative to what the algorithm would have advised, with no clearly observable peak. In general, participants tend to over-estimate apartment prices relative to the algorithm (mean of +42%, median of +24%, std.dev. of 87%), with some extreme outliers in the right tail. In the second histogram this is contrasted by the distribution of the *initial* price estimations by participants in treatments where algorithmic advice is given. While these are estimates *before* the participant received algorithmic advice for the particular task, we observe a compression of the distribution compared to estimates with no advice at all (mean of +15%, median of +6%, std.dev. of 54%). This is due to learning effects *across* tasks: algorithmic advice received for earlier tasks calibrates estimates given in later tasks.¹⁷

However, the spread across estimates is still significant. Finally, the third histogram shows the distribution of participants' final price estimations, which they make *after* receiving algorithmic advice. In contrast to the first two histograms, participants' price estimations are almost symmetrically centered around the algorithmic price prediction (mean of +7% and median of +4%), and the number of outliers is substantially lower (std.dev. of 24%). This indicates that participants respond to algorithmic advice by moving their estimation towards the advice. However, while the modal deviation from advice is zero, there is still substantial variation above and below the algorithmic price prediction.

In Table A.2 in Appendix A we show OLS regressions that predict the true price of an apartment based on the initial individual estimate of a participant and the algorithmic advice she received (with the constant being omitted). The coefficients thus represent the optimal weights participants should have given their initial estimate and the algorithmic advice in order to minimize prediction errors. While coefficients for the participant's initial estimate are significantly different from zero (and thus have informational value), they vary between 0.021 and 0.056, and the predicted coefficients for algorithmic advice are between 0.847 and 0.877, depending on algorithm and incentive condition. Thus, participants should weigh the algorithmic advice much higher than their own estimate. As Figure 2 below shows, the maximum average weight of advice across experimental conditions is less than 50%, indicating some degree of algorithm aversion.

¹⁷When we compare participants' error in the initial price estimate in their first task with the error on the initial estimate in the 10th (last) task, we observe not much change in the no advice condition. The average initial error even slightly increases from 72.7% to 77.1% ($p = 0.034$, Sign Test on matched pairs). Contrary, in the AI advice treatments, participants calibrate their initial estimates through algorithmic advice received in previous tasks for different apartments. The mean absolute error of the initial estimate decreases from 71.5% to 39.0% ($p < 0.001$) in the AI-advice condition, and from 83.0% to 40.0% in the human-AI-advice condition ($p < 0.001$).

FIGURE 1: RELATIVE DEVIATION OF INITIAL AND FINAL PRICE ESTIMATIONS FROM ALGORITHMIC PRICE PREDICTIONS

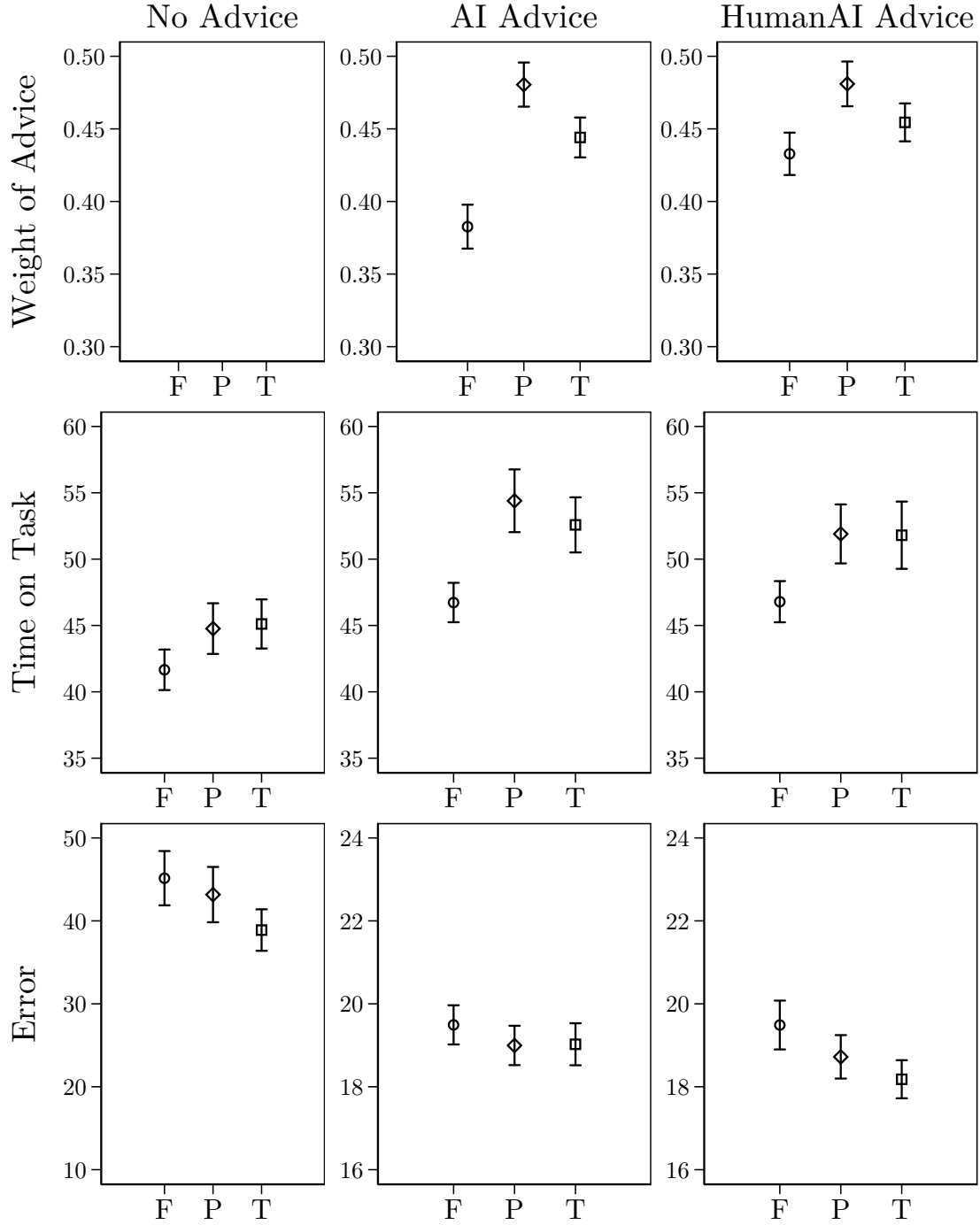


IV.B Aggregate Results

Figure 2 presents means and standard errors of our three dependent variables weight of advice, time spent on task, and performance (measured as absolute estimation error) for our fully-crossed factorial design of three incentive treatments (*fixed payment*, *performance-based incentives*, and *tournament incentives*) with three advice treatments (*no advice*, *AI advice*, and *human-AI advice*).

Concerning algorithm use, the figures in the first row show that in the *AI advice* condition, participants' *weight of advice* is substantially higher when they receive *performance pay* (mean 0.480) or compete in a *tournament* (mean 0.444) than when they are compensated with a *fixed payment* (mean 0.383). This positive effect of financial incentives on algorithm use, relative to *fixed pay*, is similar in the *human-AI advice* condition, although the margins are smaller. Further, participants whose compensation is contingent on winning a *tournament*, show a somewhat lower weight of advice than participants whose compensation is *performance-based*, but still substantially higher compared to participants who receive a *fixed payment*. Finally, for

FIGURE 2: TREATMENT AVERAGES FOR WEIGHT OF ADVICE, TIME ON THE TASK, AND ESTIMATION ERROR



Notes: F, P, and T stand for treatments ‘Fixed Payment’, ‘Performance-based incentives’ and ‘Tournament incentives’, respectively. Whiskers indicate standard errors based on OLS regression models controlling for order, apartment, and subject pool fixed effects, with robust SEs clustered by participant.

fixed-pay participants the weight of advice is visibly higher in the *human-AI* condition than in the *AI advice* condition, while there are no discernible differences in the *performance pay* and *tournament* treatments.

The second row shows how much time (in seconds) participants spend on each price estimation (“effort duration”), contingent on our treatments. Irrespective of whether they receive algorithmic advice or not, participants consistently spend more time on the price estimation task when their compensation is based on *performance* or contingent on winning a *tournament* than when they receive a *fixed payment*.¹⁸ For instance, in the *no advice* condition, *fixed pay* participants work on average for 41.66 seconds, while participants spend 44.76 seconds on the task when they receive *performance pay* and 45.12 seconds when they are competing in a *tournament*. The difference in the effort duration between unincentivized and incentivized participants in the *AI advice* and in the *human-AI advice* conditions is even bigger. However, there is no visible difference in the time spent between *performance pay* and *tournament pay*.

The last row shows participants’ performance on the task in terms of their absolute estimation error. (Note that the y-scale in the left figure in this row is different to the y-scales of the middle and right figure.) Participants in the *AI advice* and the *human-AI advice* conditions have substantially lower estimation errors compared to participants who do not receive algorithmic advice. In all three advice conditions, participants who receive *performance pay* or compete in *tournaments* have lower, or at least equally low, estimation errors as participants who receive a *fixed payment*. However, as our analysis below will show, these latter differences are generally not statistically significant at a customary level.

IV.C Regression Analyses

In Tables 2, 3, and 4, we present results from different OLS models regressing our three dependent variables weight of advice, time, and estimation error, respectively, on incentive and algorithmic advice treatment indicators. All regressions use robust standard errors clustered by participant and control for apartment, order, and subject pool fixed effects. Depending on the specification, we also include demographic control variables for gender, age, education, and nationality.¹⁹

¹⁸Differences in the time between the *no advice* treatment and the two advice treatments should be interpreted cautiously due to our one-stage vs. two-stage experimental design. See our detailed discussion below.

¹⁹We also ran all regressions reported here and below as (1) Tobit models with robust standard errors clustered by participant, (2) random (participant) effects models with robust standard errors, and (3) Tobit random effects models. All findings are identical in terms of direction and statistical significance, and thus our inferences remain unchanged. These regressions are included in our replication package. Tables A.3, A.4, and A.5 in Appendix A report results from separate OLS regression models for the first task encountered by a participant, tasks 2-5, and tasks 6-10, showcasing the robustness of our results.

Weight of Advice. Table 2 shows OLS regression results for weight of advice (based only on observations from the advice treatments). In model 1, we regress the weight of advice on a dummy variable for *performance pay* as well as a dummy variable for *tournament incentives*. In model 2, we include a dummy variable for the framing of the algorithm as *human-AI advice*. In the third model we include interaction terms of *human-AI advice* with *performance pay* as well as with *tournament incentives*. Finally, in model 4, we include all of the aforementioned dummies and interaction terms, and control for demographic characteristics. Across all models, we find that both *performance pay* as well as *tournament incentives* have a significant and positive effect on the weight of advice. We observe a difference of 0.031 in the weight of advice between *performance-based* and *tournament incentives*, which, however, is statistically not significant ($p = 0.128$).²⁰

Regressions (2), (3) and (4) indicate a positive main effect of *human-AI advice* (compared to “regular” AI advice) on advice utilization. However, as the interaction effects in model 4 indicate, this effect is mainly driven by participants with a fixed compensation contract, while the human-AI framing makes little difference for participants with performance-based incentives ($\beta = 0.000$, $p = 0.983$) or tournament incentives ($\beta = 0.011$, $p = 0.554$).²¹

Time Spent on Task. In Table 3, we present results from six OLS regression models of time spent on the experimental task. Independents include dummy variables for *performance pay* and *tournament incentives*, for *human-AI advice*, and their interaction effects. As before, regressions control for apartment, order, and subject pool fixed effects.²² The first two models use the total time spent on the estimation task (i.e., the sum of the time spent on the initial estimation task before advice and the final estimation task after receiving advice) as the dependent variable. Model 1 specifically focuses on overall incentive effects over all treatments, while model 2 only uses data from the two-stage advice treatments and also distinguishes incentive effects between *AI advice* (baseline) and *human-AI advice* conditions. The next three models only consider the time spent on the initial estimation task, before advice, either for all treatments (3), only the *no advice* treatments (4, where initial estimate time equals total time), or only the *advice* treatments (5). The last regression model 6 uses the time spent on the second estimation task (in the *advice* treatments) as dependent.

²⁰When running the regression using only data from the *AI advice* conditions, the difference is 0.035 ($p = 0.084$).

²¹Table A.6 in Appendix A reports the same regressions but with the uncensored weight of advice as dependent. Table A.7 shows that the treatment effects are robust across split-samples of multiple measures of participants’ task familiarity and overconfidence.

²²We ran the same models also with demographic controls included, reported in Table A.8 in Appendix A. Results for treatment effects reported here do not differ in any way. Some of these additional regressions indicate that male, younger, and Austrian participants as well as those who already completed a university degree tend to need less time for the experimental task.

TABLE 2: OLS REGRESSIONS OF WEIGHT OF ADVICE ON TREATMENTS

	(1)	(2)	(3)	(4)
PerformancePay	0.071*** (0.015)	0.073*** (0.015)	0.097*** (0.021)	0.095*** (0.021)
Tournament	0.041*** (0.014)	0.042*** (0.014)	0.063*** (0.020)	0.064*** (0.020)
HumanAIAdvice		0.022* (0.012)	0.051** (0.021)	0.050** (0.021)
HumanAIAdvice \times PerfPay			-0.050 (0.030)	-0.048 (0.030)
HumanAIAdvice \times Tourn			-0.040 (0.028)	-0.044 (0.028)
IsFemale				-0.011 (0.012)
Age				0.001 (0.001)
HasUnivDegree				-0.028** (0.013)
IsAustrian				-0.020 (0.013)
Constant	0.450*** (0.021)	0.438*** (0.022)	0.423*** (0.023)	0.442*** (0.041)
Observations	9,711	9,711	9,711	9,711
R-squared	0.026	0.027	0.028	0.030
N Participants	989	989	989	989

Notes: The regressions are based only on observations from the advice treatments. Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

Over all conditions (model 1), we find that participants who receive *performance pay* or compete in a *tournament* spend significantly more time in total on the price estimation task than participants who receive a *fixed payment*. However, as models 2, 4, 5, and 6 show, this effect is mainly driven by the effects of financial incentives on time spent on the initial estimate in the advice treatments, where participants form their own judgment before receiving advice. The coefficients for financial incentives for the initial estimate in the no-advice treatments (model 4), and for the second estimate (model 6), are positive but do not reach statistical significance.

As Figure 2 shows, in the *AI advice* and *human-AI advice* conditions the participants spend more total time on the price estimation tasks than when receiving no advice. This, however, is just an artifact of our two-stage design in the *advice* conditions, relative to the one stage in the

TABLE 3: OLS REGRESSIONS OF TIME SPENT ON TASK ON TREATMENTS

Dependent Conditions	Both estimates		Time spent on Initial estimate			Final estimate
	All	Advice only	All	No advice	Advice only	Advice
Model	(1)	(2)	(3)	(4)	(5)	(6)
PerformancePay	5.137*** (1.549)	7.287*** (2.793)	4.942*** (1.401)	3.205 (2.437)	6.464*** (2.379)	0.823 (0.668)
Tournament	4.854*** (1.539)	5.871** (2.543)	4.311*** (1.377)	3.641 (2.396)	5.501** (2.133)	0.370 (0.728)
HumanAIAdvice		-0.294 (2.158)			0.432 (1.814)	-0.726 (0.547)
HumanAIAdvice × PerfPay		-2.046 (3.906)			-1.999 (3.400)	-0.046 (0.845)
HumanAIAdvice × Tourn		-0.768 (3.910)			-1.729 (3.339)	0.961 (0.931)
Constant	90.163*** (2.175)	91.859*** (2.869)	73.510*** (2.002)	87.179*** (3.690)	66.331*** (2.427)	25.528*** (0.905)
Observations	14,890	9,890	14,890	5,000	9,890	9,890
R-squared	0.178	0.179	0.135	0.188	0.115	0.239
N Participants	1,489	989	1,489	500	989	989

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

no advice condition. Within advice treatments, whether the advice has a human component or not (*human-AI advice* vs. *AI advice*) has no effect on time spent on tasks.

In the regressions reported in Table A.9 in Appendix A we explore treatment effects on time spent on own estimate in more detail. To ease interpretability of treatment interactions, in model 1 we pool the performance and tournament conditions (Financial Incentives) as well as the two advice conditions (Algorithmic Advice). In model 2 we only pool the advice conditions, in model 3 only the incentive conditions, and in model 4 we report all individual treatment conditions and their interactions. The first result to point out is that the effect of advice on time spent on the initial estimate is negative in all models. That is, while the total time spent on the estimation task increases when introducing the second advice stage, the time spent on forming one’s own estimate decreases. The robust positive incentive effects in the advice treatments (evidenced in Table 3 and by post-hoc F tests at the bottom of Table A.9) can thus be understood as mitigating this negative effect of adding an advice stage on time spent on the initial estimate. In the conditions without advice, the positive effect of financial incentives is statistically weakly significant only when we pool the incentive conditions.

In sum, without advice and only one estimate to submit, financial incentives have a positive but statistically not significant effect on time spent on task. Adding an advice stage shifts some of the time spent on the own estimate to the second stage, but with financial incentives, much of this reduction in time is mitigated. Financial incentives have no effect on time spent in the second stage of the advice treatments.

TABLE 4: OLS REGRESSIONS OF ESTIMATION ERROR ON TREATMENTS

	(1)	(2)	(3)	(4)	(5)
PerformancePay	-0.183 (1.771)		-0.974 (1.613)	-1.930 (4.676)	-1.998 (4.656)
Tournament	-2.561 (1.562)		-2.683* (1.407)	-6.158 (4.126)	-6.154 (4.121)
AIAdvice		-23.225*** (1.803)	-23.191*** (1.796)	-25.510*** (3.309)	-25.486*** (3.303)
HumanAIAdvice		-23.596*** (1.810)	-23.631*** (1.804)	-25.578*** (3.317)	-25.557*** (3.313)
AIAdvice \times PerfPay				1.502 (4.765)	1.342 (4.768)
AIAdvice \times Tourn				5.568 (4.162)	5.557 (4.163)
HumanAIAdvice \times PerfPay				1.277 (4.755)	1.136 (4.750)
HumanAIAdvice \times Tourn				4.767 (4.173)	4.516 (4.201)
IsFemale = 1					-1.777 (1.419)
Age					-0.222* (0.119)
HasUnivDegree = 1					-0.175 (1.357)
IsAustrian = 1					1.482 (1.199)
Constant	18.155*** (2.700)	32.074*** (2.635)	33.345*** (3.048)	34.713*** (4.301)	40.242*** (6.028)
Observations	14,890	14,890	14,890	14,890	14,890
R-squared	0.055	0.152	0.153	0.154	0.156
N Participants	1,489	1,489	1,489	1,489	1,489

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

Estimation Error. In Table 4, we present results from five different OLS regression models mirroring the models on weight of advice in Table 2, only here with the estimation error as the dependent variable. The regression models yield highly significant main effects of *AI advice* and *human-AI advice*, which shows that participants in our algorithmic advice treatments (where they could adjust their estimate after receiving advice) perform substantially better than participants in our treatments without algorithmic advice, where participants only entered their own estimate.²³ The effects of financial incentives through *performance pay* or a *tournament* on the estimation error are negative throughout all models (implying a positive effect on performance). However, they do not reach statistical significance (with the exception of a weakly significant negative effect of *tournament incentives* in model 3).

Higher weight of advice improves performance. In Table A.10 in Appendix A we report similar regressions as the ones in Table 4, but with the weight of advice included as an explanatory variable. The results show that a higher weight of advice generally leads to a lower estimation error. One reason why we do not find higher performance in the incentive treatments despite higher advice utilization may be due to significant heterogeneity in initial estimates in our experimental task and other noise. In Appendix B we report simulations that separately reduce heterogeneity/noise in weight of advice and initial estimates across the sample, which lend some support to this explanation. Heterogeneity may also be non-random. For example, if, in line with Hypothesis 1a, financial incentives let participants put more unproductive effort into initial estimates such that initial errors are higher, but also lead to a higher weight of advice, then the net effect on eventual estimation errors may be zero. However, in our analysis in Appendix B we do not find support for such a mechanism: effects of financial incentives on initial estimate errors in the advice treatment conditions are statistically non-significant (and mostly negative).²⁴

²³In principle, we cannot rule out that the higher performance in the advice treatments is not only a result of receiving advice, but also because participants had to enter their estimate twice in our two-stage design (as opposed to once in the no-advice treatments). However, we see no reason to assume that a rational decision-maker would alter and improve her estimate without new information, only because being asked twice.

²⁴We test whether our results reported in Tables 2, 3, and 4 are robust to using a Bonferroni correction that adjusts significance thresholds for all possible pairwise comparisons between treatments. This implies dividing alpha-levels by 15 for Weight of Advice, by 36 for Time Spent on Task, and by 36 for Estimation Error. If we do this for the Weight of Advice regressions reported in Table 2, then the positive effect of performance-based and tournament incentives on the weight of advice remains unchanged, but the difference between the AI and the Human-AI condition is no longer significant. Dividing alpha-levels by 36 for the Time Spent on Task regressions in Table 3 implies that the positive effect of incentives on time in the advice conditions in columns (2) and (5) are no longer significant. However, the findings in columns (1) and (3) remain unchanged; when collapsing the advice and no advice conditions, the positive effect of incentives on time is still significant. Applying the same correction to Estimation Error regressions in Table 4, the findings remain largely unchanged: Participants in the (two-staged) algorithmic advice treatments still perform better than participants in the (one-staged) no-advice treatments.

IV.D Additional Insights from the Post-Experiment Questionnaire

Our post-experimental questionnaire obtained information on participants’ attitudes and perceptions of the experimental tasks, allowing for some sanity checks of our results and explorative analysis of mechanisms. In Table 5, we display results from regressing the weight of advice on items from the post-experiment questionnaire and participants’ demographic characteristics, while controlling for experimental conditions as well as apartment, order, and subject pool fixed effects.

In model 1 of Table 5, we include six items that measure how familiar participants are with the experimental task of estimating prices of Airbnb apartments in Vienna. We find that individuals who have previously used the online platform Airbnb to book an apartment have a significantly lower weight of advice, while self-reported unfamiliarity with the price estimation task²⁵ is positively related to participants’ advice utilization. Self-assessed knowledge of Vienna, residency in Vienna, and previous use of an AI algorithm in a similar task are not associated with participants’ weight of advice.

Model 2 focuses on three personality characteristics of participants. We find that task enjoyment as well as the willingness to take risks are unrelated to advice utilization. Interestingly, we obtain a significant negative effect of overconfidence on the weight of advice. The binary indicator *overconfidence* equals 1 if a participant overestimated his or her own ability to correctly estimate prices of Airbnb apartments (i.e., the actual mean absolute error is greater than the expected mean absolute error). In model 3 we include questionnaire items from models 1 and 2 simultaneously and all of our inferences remain unchanged.

In model 4 we find a significant positive association between participants’ perception of the algorithm as a credible source of advice in the price estimation task and their weight of advice. Furthermore, we observe a significant positive relation between participants’ self-stated use of algorithmic advice and the measured weight of advice, as well as a significant negative relation between their self-stated use of contextual information and the weight of advice. These results on self-stated variables serve as sanity checks for our results on observational variables.²⁶

In Table 6 we report regression results where we use the questionnaire items not as predictors of weight of advice but rather as dependent variables. In column 1, we regress task enjoyment on our treatment dummies for *performance pay*, *tournament*, *AI advice*, and *human-AI advice*, while controlling for subject pool fixed effects. Our findings indicate that incentivized participants perceive the task as significantly less enjoyable than unincentivized participants who

²⁵The construct “unfamiliarity” is based on three items that ask participants how (un)familiar they are with the price setting of Airbnb apartments in Vienna.

²⁶In Table A.7 in Appendix A, we report results from OLS regressions which show that our results on treatment effects on the weight of advice are robust across split-samples of multiple measures of participants’ task familiarity and overconfidence. These results also indicate that our findings are not driven by task complexity.

TABLE 5: OLS REGRESSIONS OF WEIGHT OF ADVICE ON QUESTIONNAIRE ITEMS

	(1)	(2)	(3)	(4)
IsFemale	-0.013 (0.012)	-0.015 (0.012)	-0.017 (0.012)	-0.006 (0.011)
Age	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)	0.002 (0.001)
HasUnivDegree	-0.022* (0.013)	-0.024* (0.013)	-0.017 (0.013)	-0.017 (0.012)
IsAustrian	-0.021 (0.013)	-0.027** (0.012)	-0.025** (0.013)	-0.007 (0.011)
Resident of Vienna [0/1]	0.013 (0.024)		0.015 (0.024)	
Self-assessed Vienna knowledge [0-10]	-0.002 (0.004)		-0.003 (0.004)	
Has used Airbnb [0/1]	-0.055*** (0.015)		-0.054*** (0.014)	
× Has used Airbnb in Vienna [0/1]	0.001 (0.017)		0.003 (0.017)	
Unfamiliarity with task [0-10, avg. of 3 items]	0.011*** (0.003)		0.010*** (0.003)	
Previously used AI in similar task [0/1]	0.022 (0.018)		0.012 (0.018)	
Task enjoyment [0-10]		-0.003 (0.003)	0.001 (0.003)	
Willingness to take risks [0-10]		0.003 (0.003)	0.004 (0.003)	
Overconfident (real error > exp. error) [0/1]		-0.069*** (0.013)	-0.064*** (0.013)	
AI is credible source for task [0-10]				0.013*** (0.003)
Self-stated use of alg. advice [0-100%]				0.003*** (0.000)
Self-stated use of context info [0-100%]				-0.002*** (0.000)
Constant	0.446*** (0.055)	0.507*** (0.051)	0.466*** (0.063)	0.386*** (0.050)
Observations	9,711	9,711	9,711	9,711
R-squared	0.041	0.038	0.049	0.085
N Participants	989	989	989	989

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

TABLE 6: OLS REGRESSIONS OF QUESTIONNAIRE ANSWERS ON TREATMENTS

	(1) Task enjoyment	(2) Algorithm is Credible	(3) Use of algorithm	(4) Use of context info
PerformancePay	-0.445*** (0.121)	0.206 (0.139)	7.779*** (1.823)	-0.591 (1.215)
Tournament	-0.376*** (0.118)	0.291** (0.137)	6.988*** (1.741)	0.443 (1.176)
AIAdvice	0.662*** (0.124)			
HumanAIAdvice	0.755*** (0.125)	0.219* (0.112)	2.900** (1.455)	-0.399 (0.986)
Constant	7.505*** (0.148)	6.024*** (0.193)	47.086*** (2.200)	82.298*** (1.521)
Observations	1,489	989	989	989
R-squared	0.050	0.009	0.029	0.001

Notes: Robust standard errors in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for subject pool fixed effects.

receive a fixed payment. Presumably, the incentive-induced pressure to perform well undermines individuals' intrinsic task motivation. Further, participants who receive advice from an AI or a human-AI framed algorithm enjoy the task of estimating apartment prices more than individuals who do not receive advice. An explanation could be that by receiving algorithmic advice, the task is perceived to be easier and less strenuous, increasing participants' intrinsic task motivation.

The regression reported in model 2 relates the perceived credibility of the algorithm to our treatment conditions. Participants who receive incentives perceive the algorithm to be a more credible source of advice than participants who receive a fixed payment, but the effect is significant only for the tournament condition. When the algorithm is framed as *human-AI advice*, the perceived source credibility is weakly significantly higher than when the algorithm is framed as *AI advice*.²⁷

In model 3 we regress participants' self-reported use of the algorithm on our incentive and algorithm framing treatments. In line with our results on observed weight of advice, we find that *performance pay* and *tournament incentives* are significantly and positively related to the

²⁷This effect remains positive but becomes non-significant when running this model separately for the three compensation contract treatments.

stated extent of algorithm use. This provides corroborating evidence that financial incentives have a positive effect on algorithmic advice utilization. We also observe that participants in the *human-AI advice* condition self-report to rely significantly more on the algorithm than participants in the *AI advice* condition. In model 4 we regress participants’ self-reported use of contextual information on treatment indicators but do not observe significant effects.²⁸

IV.E Discussion

Competing Hypotheses 1a and 1b. Our first set of hypotheses posits the tension between (1a) previous research suggesting that financial incentives “backfire” in the presence of algorithmic decision-aids due to increased but unproductive individual effort, and (1b) economic-theory-based reasoning that financial incentives translate into more use of a sufficiently accurate algorithm.

As described in our discussions of Table 2, we find evidence that individuals who are compensated with performance-based or tournament incentives rely significantly more on the algorithmic advice in our task than individuals who receive a fixed payment, providing support for Hypothesis 1b rather than Hypothesis 1a. The explanation for Hypothesis 1a includes increased individual effort, while a prediction of standard utility theory is ambiguous in our experimental setup: financial incentives may increase effort into individual (first-stage) estimates but may reduce effort in advised (second-stage) estimates because the decision-maker would follow the advice instead of doing (more) effortful analyses. Our finding that overall effort duration increases with financial incentives, mainly due to increased time in the initial estimate formation, is consistent with both mechanisms.

In terms of performance, we observe that individuals who receive tournament incentives make better price estimations than individuals who receive a fixed payment, but these effects are not statistically significant in most cases. However, our results are clearly not in line with Hypothesis 1a, which predicted a backfiring effect of financial incentives on performance.

That said, incentives increase the time spent on the initial estimate and increase the weight of beneficial algorithmic advice, but do not result in significant improvements of performance. In terms of time spent, performance effects may be specific to the task. Namely, the nature of our prediction task (without feedback) may be that task-relevant skills cannot be improved by spending more time on the task. The question why – with an advantageous AI prediction aid – a (significantly) higher weight of advice in treatments with financial incentives does

²⁸In Table A.11 in Appendix A we run these four models with treatment interaction terms. The only additional insights gained are that the higher rating of task enjoyment under (non-humanised) AI Advice seems to be mainly driven by observations in the tournament condition, and that the weakly significant effect of Human-AI Advice (vs regular AI advice) on perceived credibility is statistically not significant anymore when splitting up by treatments, as mentioned in the previous note.

not directly translate into lower prediction error, is not straightforward to answer. The final prediction error is not only subject to the weight of advice but also to the initial estimate performance as well as other idiosyncratic effects when submitting the second estimate. It appears that while heterogeneity in weight of advice is small enough to identify treatment effects of financial incentives, significant heterogeneity in initial estimates in our experimental task and other noise obfuscate a translation to smaller estimation errors. Simulations that reduce heterogeneity/noise in weight of advice and initial estimates across the sample lend some support to this conjecture (Appendix B).

Thus, our experimental findings provide a more nuanced picture on how incentives influence the interaction between human decision-makers and algorithms. While we find support for the presumption that incentives trigger decision-makers to exert more effort (i.e., they work longer), we do not observe that this additional effort duration undermines algorithm use. Rather, financial incentives lead to a higher reliance on algorithmic advice, and positive but statistically not significant effects on performance. Overall, our results indicate that financial incentives do not hinder decision-making with algorithmic advice. However, while effort costs for participants are relatively low in our setting, they may be higher in practice. High costs of time or effort may mitigate the effects of incentives on time spent on the individual estimate. At the same time, high effort costs together with an increase of time spent and low or zero performance effects may yield overall negative effects of incentives on efficiency.

Hypothesis 2. Our second hypothesis postulates that decision-makers with tournament incentives rely less on algorithmic advice and as a consequence perform worse than decision-makers with performance-based incentives. Although our empirical results are directionally consistent with this prediction in terms of the weight of advice, the differences between performance-based and tournament incentives are rather small and typically do not reach statistical significance.²⁹ Both types of financial incentives lead to consistently higher algorithm use, greater effort duration, and no performance drop, compared to fixed payments. Therefore, we conclude that both, performance-based and tournament incentives, have positive effects in decision-making tasks with algorithmic advice.

Hypothesis 3. Based on a contemporaneous stream of research on algorithm aversion, which suggests that decision-makers prefer to receive advice from human experts rather than algo-

²⁹One reason why we only find small differences between effects of individual performance-based and tournament incentives could be that, in order to keep stake sizes and payoff ranges similar, we used an arguably rather weak form of tournament contracts (one loser and one winner in each pair). Any differences may have been more pronounced if we had used a tournament design with higher competition and a larger spread of payoffs (e.g., the best performing individual out of all participants receives a large reward of EUR 1,000).

rithms, we hypothesize that reliance on algorithmic advice is higher when an algorithm is framed as also considering human expertise. The results of our experiment provide some support for this hypothesis. Individuals’ weight of advice is higher when the algorithm is presented as human-augmented. The positive effect of the human-AI framing on advice utilization is particularly strong for individuals with fixed-pay contracts, while it has little effect on individuals with performance-based or tournament contracts. Findings from the post-experiment questionnaire corroborate this conclusion, in that the perceived credibility of the human-AI framed advice is higher compared to the AI advice. Thus, we conclude that subtle changes in the framing of an otherwise identical algorithm may have substantial effects on acceptance of algorithmic advice, especially for decision-makers who are not exposed to more direct monetary incentives.

V CONCLUSION

We investigate how the design of decision-makers’ compensation contracts and the framing of an artificial intelligence algorithm influence advice utilization, exerted effort duration, and eventual performance. Results from our experimental study show that, even though algorithm aversion may exist in some task settings, financial incentives do not generally “backfire” in terms of how they motivate decision-makers to consider (augmented) algorithmic advice in their judgments. In our contemporary setting with a contextually-rich price estimation task, we find that decision-makers with performance-based or tournament incentives rely significantly more on algorithmic advice than decision-makers who receive a fixed payment. We also observe that the use of the algorithmic decision aid is stronger when the involvement of human experts in the development of the AI algorithm is highlighted to decision-makers, particularly for decision-makers with fixed-pay contracts. Thereby, our study contributes not only to concurrent academic literature on algorithm aversion in disciplines such as accounting, management, economics, and psychology, but also to managerial practice. Organizations who want to encourage AI use of their employees can provide them with financial performance incentives. Where such incentives are practically not feasible, organizations can increase acceptance of AI tools by highlighting the involvement of human expert knowledge.

Our results contrast prominent earlier findings of a backfiring effect of financial incentives in presence of an algorithmic decision aid (in particular Ashton, 1990; Awasthi and Pratt, 1990). Our study constitutes a conceptual replication rather than a direct one. For example, Ashton (1990) used a different task (predicting Moody’s bond ratings based on three financial ratios as input), the algorithmic decision-aid was simpler (a simple weighing of the three financial ratios), and participants were 182 employees of an auditing firm. Awasthi and Pratt (1990) had 70 students work on three accounting-related problems. We use a different, more modern price prediction task, employ a more sophisticated algorithm, and draw on a large sample of almost

1500 students with heterogeneous backgrounds. In addition, our experiment took place about 30 years later, and attitudes towards algorithmic advice may have significantly changed over time. Thus, our main message with respect to this older literature is that the still prominently cited backfiring effect may not hold anymore.

As every study, our experiment is subject to limitations, which provide ample opportunities for future research. First, we use a specific context-rich task in which participants have to estimate the prices of Airbnb apartments. In light of research showing that algorithm aversion is a task-dependent phenomenon (Castelo et al., 2019; Hertz and Wiese, 2019), it could well be the case that our findings are not generalizable to other, more subjective tasks used in previous studies, such as detecting emotions, predicting the funniness of jokes, or providing dating advice. Second, to the best of our knowledge, we are also one of the first studies that directly compares decision-makers' reliance on AI advice with human-AI combined advice instead of the typically used human-expert advice. Although we would expect directionally similar results when using purely human expert advice, these effects remain an open empirical question to explore. Third, as most laboratory experiments, we rely on student participants. Although we use a broad sample of undergraduate and postgraduate students from three different universities, one could argue that older and more experienced individuals may behave differently when it comes to using advice from algorithmic decision aids. However, it is less clear why treatment effects would be different for a different subject population. Fourth, our experiment intentionally abstracts away from many strategic considerations that typically arise in organizational contexts. For instance, decision-makers within firms could expect that by relying on algorithmic advice instead of making own judgments, it becomes more likely that they might eventually be fully replaced by an algorithm. Such a dynamic perspective on human-algorithm interactions may be a fruitful field of future research.

Finally, in line with our expectations, we find that decision-makers who receive a fixed compensation have a higher weight of advice when the algorithm is framed as including also human expert judgment than when it is framed as a pure AI algorithm. At the same time, we observe no significant uptake in advice utilization from such human framing of the algorithmic advice when decision-makers are compensated with performance-based or tournament incentives. A potential explanation could be that financial incentives make decision-makers act more rationally and thus they are less influenced by the framing of the algorithmic advice as long as the accuracy of the algorithmic prediction is identical. We encourage future research to explore these questions.

REFERENCES

- Allen, R. T. and Choudhury, P. (2022). Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organization Science*, 33(1):149–169.
- Arkes, H. R., Dawes, R. M., and Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, 37(1):93–110.
- Ashton, R. H. (1990). Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback, and justification. *Journal of Accounting Research*, 28:148–180.
- Awasthi, V. and Pratt, J. (1990). The effects of monetary incentives on effort and decision performance: The role of cognitive characteristics. *The Accounting Review*, 65(4):797–811.
- Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, 126(4):1472–1503.
- Bauer, K., Hinz, O., van der Aalst, W., and Weinhardt, C. (2021). Expl(AI)n it to me – explainable AI and information systems research. *Business & Information Systems Engineering*, 63:79–82.
- Bauer, K., von Zahn, M., and Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users’ information processing. *Information Systems Research*, 34(4):1582–1602.
- Bonaccio, S. and Dalal, R. S. (2006). 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.
- Bonner, S. E., Hastie, R., Sprinkle, G. B., and Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research*, 12(1):19–64.
- Bonner, S. E. and Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society*, 27(4):303–345.
- Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.
- Camerer, C. F. and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1):7–42.
- Cao, T., Duh, R.-R., Tan, H.-T., and Xu, T. (2022). Enhancing auditors’ reliance on data analytics under inspection risk using fixed and growth mindsets. *The Accounting Review*, 97(3):131–153.
- Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.
- Charness, G., Gneezy, U., and Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, 131:141–150.
- Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256.
- Chen, C. X., Hudgins, R., and Wright, W. F. (2022). The effect of advice valence on the perceived credibility of data analytics. *Journal of Management Accounting Research*, 34:97–116.
- Choudhury, P., Starr, E., and Agarwal, R. (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, 41(8):1381–1411.
- Commerford, B. P., Dennis, S. A., Joe, J. R., and Ulla, J. W. (2022). Man versus machine: Complex estimates and auditor reliance on artificial intelligence. *Journal of Accounting Research*, 60(1):171–201.
- Costello, A. M., Down, A. K., and Mehta, M. N. (2020). Machine + man: A field experiment on the role of discretion in augmenting AI-based lending models. *Journal of Accounting and Economics*, 70(2):101360.
- Dargnies, M. P., Hakimov, R., and Kübler, D. (forthcoming). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*.
- Dell’Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraymer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper 24-013.
- Dellermann, D., Ebel, P., Söllner, M., and Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61:637–643.
- Dietvorst, B. J. and Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, 31(10):1302–1314.

- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology General*, 144(1):114–126.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.
- Ding, S. and Beaulieu, P. (2011). The role of financial incentives in balanced scorecard-based performance evaluations: Correcting mood congruency biases. *Journal of Accounting Research*, 49(5):1223–1247.
- Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.
- Efendić, E., Van de Calseyde, P. P. F. M., and Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157:103–114.
- Emett, S. A., Kaplan, S. E., Mauldin, E., and Pickerd, J. S. (2023). Auditing with data and analytics: External reviewers’ judgments of audit quality and effort. *Contemporary Accounting Research*, 40(4):2314–2339.
- Estep, C., Griffith, E. E., and MacKenzie, N. L. (2023). How do financial executives respond to the use of artificial intelligence in financial reporting and auditing? *Review of Accounting Studies*.
- Farrell, A. M., Goh, J. O., and White, B. J. (2014). The effect of performance-based incentive contracts on system 1 and system 2 processing in affective decision contexts: fMRI and behavioral evidence. *The Accounting Review*, 89(6):1979–2010.
- Frechette, G. R. (2015). Laboratory Experiments: Professionals Versus Students. In Frechette, G. R. and Schotter, A., editors, *Handbook of Experimental Economic Methodology*, pages 360–390. Oxford University Press.
- Gaessler, F. and Piezunka, H. (2023). Training with AI: Evidence from chess computers. *Strategic Management Journal*, 44(11):2724–2750.
- Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.
- Goldstein, I. M., Lawrence, J., and Miner, A. S. (2017). Human-machine collaboration in cancer and beyond: The centaur care model. *JAMA Oncology*, 3(10):1303–1304.
- Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.
- Harvey, N. and Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, 70(2):117–133.
- Hertz, N. and Wiese, E. (2019). Good advice is beyond all price, but what if it comes from a machine? *Journal of Experimental Psychology: Applied*, 25(3):386–395.
- Hoffman, M., Kahn, L. B., and Li, D. (2017). Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800.
- Hossain, T. and Okui, R. (2013). The binarized scoring rule. *The Review of Economic Studies*, 80(3):984–1001.
- Jung, M. and Seiter, M. (2021). Towards a better understanding on mitigating algorithm aversion in forecasting: An experimental study. *Journal of Management Control*, 32(4):495–516.
- Kawaguchi, K. (2020). When will workers follow an algorithm? A field experiment with a retail business. *Management Science*, 67(3):1670–1695.
- Keding, C. and Meissner, P. (2021). Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. *Technological Forecasting and Social Change*, 171:120970.
- Kellogg, K. C., Valentine, M. A., and Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1):366–410.
- Krakovski, S., Luger, J., and Raisch, S. (2023). Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal*, 44(6):1425–1452.
- Labro, E., Lang, M., and Omartian, J. D. (2023). Predictive analytics and centralization of authority. *Journal of Accounting and Economics*, 75(1):101526.
- Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864.
- Libby, R. and Lipe, M. G. (1992). Incentives, effort, and the cognitive processes involved in accounting-related judgments. *Journal of Accounting Research*, 30(2):249–273.

- Libby, R. and Luft, J. (1993). Determinants of judgment performance in accounting settings: Ability, knowledge, motivation, and environment. *Accounting, Organizations and Society*, 18(5):425–450.
- Libby, R. and Witz, P. D. (2024). Can artificial intelligence reduce the effect of independence conflicts on audit firm liability? *Contemporary Accounting Research*, 41(2):1346–1375.
- Liu, M. (2022). Assessing human information processing in lending decisions: A machine learning approach. *Journal of Accounting Research*, 60(2):607–651.
- Liu, M., Tang, X., Xia, S., Zhang, S., Zhu, Y., and Meng, Q. (forthcoming). Algorithm aversion: Evidence from ridesharing drivers. *Management Science*.
- Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.
- March, C., Ziegelmeyer, A., Greiner, B., and Cyranek, R. (2016). Pay few subjects but pay them well: Cost-effectiveness of random incentive systems. *SSRN Working Paper*.
- Murray, A., Rhymer, J., and Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, 46(3):552–571.
- Neumann, M., Hengeveld, M., Niessen, A. S. M., Tendeiro, J. N., and Meijer, R. R. (2022). Education increases decision-rule use: An investigation of education and incentives to improve decision making. *Journal of Experimental Psychology: Applied*, 28(1):166–178.
- Ottaviani, M. and Sørensen, P. N. (2006). The strategy of professional forecasting. *Journal of Financial Economics*, 81(2):441–466.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, Yokohama, Japan, Article No. 237.
- Prahl, A. and Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702.
- Raisch, S. and Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1):192–210.
- Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1):31–38.
- Samuels, J. A. and Whitecotton, S. M. (2011). An effort based analysis of the paradoxical effects of incentives on decision-aided performance. *Journal of Behavioral Decision Making*, 24(4):345–360.
- Schlag, K. H., Tremewan, J., and van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18(3):457–490.
- Schotter, A. (2023). *Advice, Social Learning and the Evolution of Conventions*. Cambridge University Press, Cambridge, MA.
- Schotter, A. and Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review of Economics*, 6(1):103–128.
- Shrestha, Y. R., Ben-Menahem, S. M., and von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(3):66–83.
- Sprinkle, G. B. (2000). The effect of incentive contracts on learning and performance. *The Accounting Review*, 75(3):299–326.
- Tasoff, J. and Zhang, W. (2022). The performance of time-preference and risk-preference measures in surveys. *Management Science*, 68(2):1149–1173.
- Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234.
- Wang, W., Gao, G., and Agarwal, R. (forthcoming). Friend or foe? Teaming between artificial intelligence and workers with variation in experience. *Management Science*.
- Wilson, H. J. and Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, 96(4):114–123.
- Zellner, M., Abbas, A. E., Budescu, D. V., and Galstyan, A. (2021). A survey of human judgement and quantitative forecasting methods. *Royal Society Open Science*, 8(2):201187.

APPENDIX

A ADDITIONAL TABLES AND FIGURES

TABLE A.1: PARTICIPANT ATTRIBUTES ACROSS TREATMENTS

	No Advice	AI Advice	Human-AI Advice
Fixed Pay			
Avg. Age	24.80	25.35	24.47
% Female	62.27	58.86	64.51
% Univ.Degree	58.58	58.68	54.83
% Austrian	54.43	52.10	47.85
N	169	167	186
Performance Pay			
Avg. Age	25.31	25.11	24.40
% Female	58.14	56.47	62.67
% Univ.Degree	54.65	47.65	48.59
% Austrian	59.30	58.82	54.93
N	172	170	156
Tournament Pay			
Avg. Age	24.83	24.69	24.20
% Female	57.86	64.88	58.97
% Univ.Degree	56.60	60.71	44.23
% Austrian	45.91	53.57	54.49
N	159	168	142

TABLE A.2: OLS REGRESSIONS PREDICTING THE TRUE PRICE BASED ON PARTICIPANT'S INITIAL ESTIMATE AND PROVIDED ALGORITHMIC ADVICE

	No Advice		AI Advice		HumanAI advice	
Initial Estimate	0.026*** (0.005)	0.021*** (0.007)	0.053*** (0.009)	0.056*** (0.010)	0.050*** (0.009)	0.046*** (0.010)
IniE \times PerfPay		0.005 (0.011)		0.020 (0.018)		0.031 (0.021)
IniE \times Tournament		0.013 (0.010)		-0.018 (0.018)		-0.004 (0.019)
Alg. Advice	0.871*** (0.006)	0.877*** (0.010)	0.849*** (0.010)	0.847*** (0.010)	0.853*** (0.009)	0.856*** (0.011)
AlgA \times PerfPay		-0.006 (0.014)		-0.022 (0.019)		-0.034 (0.023)
AlgA \times Tournament		-0.015 (0.013)		0.019 (0.019)		0.005 (0.020)
Observations	5,000	5,000	5,050	5,050	4,840	4,840
R-squared	0.937	0.937	0.938	0.938	0.938	0.938
N Participants	500	500	505	505	484	484

Notes: The dependent in all regressions is the true price of the apartment, and the constant is omitted. Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

TABLE A.3: OLS REGRESSIONS OF WEIGHT OF ADVICE ACROSS DIFFERENT PERIODS

	Task 1		Tasks 2-5		Tasks 6-10	
	(1)	(2)	(3)	(4)	(5)	(6)
PerformancePay	0.143*** (0.035)	0.136*** (0.035)	0.093*** (0.024)	0.090*** (0.024)	0.090*** (0.024)	0.090*** (0.024)
Tournament	0.045 (0.035)	0.048 (0.035)	0.057** (0.024)	0.059** (0.024)	0.070*** (0.023)	0.071*** (0.023)
HumanAIAdvice	0.089*** (0.034)	0.089*** (0.034)	0.039 (0.024)	0.039 (0.024)	0.052** (0.024)	0.050** (0.024)
HumanAIAdvice \times PerfPay	-0.111** (0.049)	-0.109** (0.049)	-0.045 (0.035)	-0.044 (0.035)	-0.042 (0.034)	-0.041 (0.034)
HumanAIAdvice \times Tourn	-0.049 (0.048)	-0.058 (0.048)	-0.026 (0.033)	-0.032 (0.033)	-0.051 (0.033)	-0.051 (0.033)
IsFemale		-0.017 (0.021)		-0.015 (0.014)		-0.005 (0.014)
Age		0.001 (0.002)		0.001 (0.002)		0.000 (0.002)
HasUnivDegree		-0.056** (0.022)		-0.037** (0.016)		-0.016 (0.015)
IsAustrian		0.009 (0.021)		-0.006 (0.015)		-0.034** (0.015)
Constant	0.420*** (0.043)	0.422*** (0.069)	0.358*** (0.030)	0.359*** (0.049)	0.303*** (0.027)	0.333*** (0.047)
Observations	972	972	3,886	3,886	4,853	4,853
R-squared	0.033	0.041	0.012	0.015	0.038	0.040
N Participants	972	972	989	989	989	989

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects. Models 1 and 2 have fewer than 989 participants since for some of them WOA is not defined (initial estimate = algorithmic advice).

TABLE A.4: OLS REGRESSIONS OF TIME SPENT ON TASK ACROSS DIFFERENT PERIODS

	Task 1		Tasks 2-5		Tasks 6-10	
	All	Advice only	All	Advice only	All	Advice Only
PerformancePay	15.428*** (3.706)	24.012*** (6.920)	5.195*** (1.791)	8.136** (3.272)	3.049** (1.369)	3.552 (2.357)
Tournament	13.451*** (3.823)	20.327*** (7.699)	4.163*** (1.598)	4.925* (2.745)	3.750** (1.493)	4.020* (2.324)
HumanAIAdvice		4.440 (5.079)		-1.579 (2.262)		0.015 (2.150)
HumanAIAdvice \times PerfPay		-21.442** (9.398)		-1.915 (4.551)		1.156 (3.426)
HumanAIAdvice \times Tourn		-9.780 (10.200)		-0.239 (3.984)		0.128 (3.799)
Constant	77.588*** (5.595)	72.329*** (7.455)	45.909*** (2.061)	47.650*** (2.785)	35.644*** (1.777)	35.369*** (2.392)
Observations	1,489	989	5,956	3,956	7,445	4,945
R-squared	0.032	0.049	0.014	0.018	0.015	0.020
N Participants	1,489	989	1,489	989	1,489	989

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

TABLE A.5: OLS REGRESSIONS OF ESTIMATION ERROR ACROSS DIFFERENT PERIODS

	Task 1		Tasks 2-5		Tasks 6-10	
	(1)	(2)	(3)	(4)	(5)	(6)
PerformancePay	-1.703 (2.522)	-0.668 (6.939)	-2.112 (1.801)	-5.710 (5.173)	0.063 (1.587)	0.602 (4.563)
Tournament	-4.125* (2.438)	-5.675 (6.382)	-4.196** (1.704)	-10.417** (4.941)	-1.241 (1.285)	-2.914 (3.754)
AIAdvice	-17.672*** (2.707)	-17.006*** (5.555)	-23.348*** (1.954)	-28.861*** (4.330)	-24.069*** (1.809)	-24.403*** (2.644)
HumanAIAdvice	-18.858*** (2.682)	-19.768*** (5.436)	-23.514*** (2.004)	-27.425*** (4.358)	-24.631*** (1.786)	-25.258*** (2.638)
AIAdvice \times PerfPay		-2.945 (7.540)		5.812 (5.360)		-1.313 (4.641)
AIAdvice \times Tourn		1.066 (6.884)		10.709** (5.005)		2.317 (3.820)
HumanAIAdvice \times PerfPay		-0.726 (7.352)		4.316 (5.355)		-0.854 (4.622)
HumanAIAdvice \times Tourn		2.976 (6.764)		7.644 (5.073)		2.398 (3.828)
IsFemale		-4.314* (2.339)		-1.888 (1.524)		-1.175 (1.420)
Age		-0.316 (0.210)		-0.176 (0.145)		-0.238** (0.112)
HasUnivDegree		-1.261 (2.205)		0.255 (1.509)		-0.347 (1.346)
IsAustrian		-0.956 (2.045)		2.409* (1.378)		1.268 (1.147)
Constant	30.640*** (4.609)	42.169*** (10.593)	33.477*** (4.133)	40.378*** (7.999)	30.634*** (1.934)	36.675*** (3.930)
Observations	1,489	1,489	5,956	5,956	7,445	7,445
R-squared	0.115	0.120	0.156	0.161	0.166	0.168
N Participants	1,489	1,489	1,489	1,489	1,489	1,489

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

TABLE A.6: OLS REGRESSIONS OF UNCENSORED WEIGHT OF ADVICE
ON TREATMENTS

	(1)	(2)	(3)	(4)
PerformancePay	0.104*** (0.026)	0.107*** (0.026)	0.112*** (0.032)	0.112*** (0.032)
Tournament	0.072*** (0.023)	0.074*** (0.023)	0.109*** (0.032)	0.111*** (0.031)
HumanAIAdvice		0.040** (0.020)	0.066** (0.032)	0.065** (0.031)
HumanAIAdvice \times PerfPay			-0.008 (0.052)	-0.007 (0.052)
HumanAIAdvice \times Tourn			-0.071 (0.045)	-0.074* (0.044)
IsFemale				0.006 (0.020)
Age				0.003 (0.003)
HasUnivDegree				-0.034 (0.021)
IsAustrian				-0.036* (0.022)
Constant	0.588*** (0.044)	0.567*** (0.044)	0.553*** (0.046)	0.522*** (0.068)
Observations	9,711	9,711	9,711	9,711
R-squared	0.012	0.013	0.014	0.015
N Participants	989	989	989	989

Notes: Dependent is the weight of advice, but not censored between 0 and 1. Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

TABLE A.7: OLS REGRESSIONS OF WEIGHT OF ADVICE ACROSS SPLIT-SAMPLES RELATED TO UNFAMILIARITY AND OVERCONFIDENCE

	Vienna knowledge		Airbnb experience		Task familiarity		Overconfident	
	Low	High	Low	High	Low	High	No	Yes
PerformancePay	0.097*** (0.027)	0.082** (0.034)	0.142*** (0.045)	0.080*** (0.024)	0.061** (0.027)	0.126*** (0.035)	0.148*** (0.038)	0.081*** (0.026)
Tournament	0.051** (0.026)	0.083** (0.032)	0.079* (0.042)	0.050** (0.023)	0.022 (0.023)	0.126*** (0.036)	0.053 (0.035)	0.078*** (0.025)
HumanAIAdvice	0.049* (0.026)	0.042 (0.034)	0.047 (0.043)	0.049** (0.024)	0.014 (0.026)	0.100*** (0.034)	0.056* (0.032)	0.047* (0.026)
HumanAIAdvice \times PerfPay	-0.066* (0.037)	-0.002 (0.050)	-0.112* (0.063)	-0.030 (0.034)	0.002 (0.037)	-0.112** (0.048)	-0.085 (0.054)	-0.033 (0.036)
HumanAIAdvice \times Tourn	-0.026 (0.035)	-0.068 (0.046)	-0.059 (0.058)	-0.032 (0.032)	0.016 (0.034)	-0.117** (0.047)	-0.067 (0.048)	-0.036 (0.034)
IsFemale	0.000 (0.015)	-0.032 (0.021)	0.006 (0.026)	-0.018 (0.014)	-0.028* (0.015)	0.012 (0.020)	-0.032 (0.021)	-0.007 (0.014)
Age	0.000 (0.002)	0.001 (0.002)	-0.002 (0.002)	0.001 (0.002)	0.002 (0.002)	0.000 (0.003)	0.004 (0.003)	0.001 (0.002)
HasUnivDegree	-0.019 (0.016)	-0.038* (0.023)	-0.030 (0.028)	-0.019 (0.015)	-0.025 (0.016)	-0.033 (0.021)	-0.044* (0.025)	-0.020 (0.016)
IsAustrian	-0.011 (0.016)	-0.026 (0.021)	-0.006 (0.026)	-0.029** (0.014)	-0.034** (0.015)	0.007 (0.021)	-0.035 (0.022)	-0.019 (0.015)
Constant	0.449*** (0.052)	0.425*** (0.070)	0.524*** (0.070)	0.420*** (0.049)	0.431*** (0.049)	0.449*** (0.075)	0.458*** (0.077)	0.405*** (0.049)
Observations	6,546	3,165	2,370	7,341	5,512	4,199	3,422	6,289
R-squared	0.029	0.039	0.040	0.031	0.031	0.040	0.037	0.035
N Participants	666	323	241	748	562	427	348	641

Notes: Dependent in all regressions is weight of advice. Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

TABLE A.8: OLS REGRESSIONS OF TIME SPENT ON TASK, WITH DEMOGRAPHICS

Dependent Conditions	Time spent on					
	Both estimates		Initial estimate			Final estimate
	All	Advice only	All	No advice	Advice only	Advice
Model	(1)	(2)	(3)	(4)	(5)	(6)
PerformancePay	5.224*** (1.566)	7.274*** (2.792)	5.036*** (1.417)	3.551 (2.454)	6.400*** (2.374)	0.874 (0.672)
Tournament	4.861*** (1.538)	6.040** (2.548)	4.325*** (1.376)	3.638 (2.383)	5.610*** (2.134)	0.430 (0.732)
HumanAIAdvice		-0.361 (2.170)			0.361 (1.826)	-0.722 (0.543)
HumanAIAdvice × PerfPay		-1.925 (3.880)			-1.906 (3.380)	-0.019 (0.837)
HumanAIAdvice × Tourn		-0.940 (3.959)			-1.881 (3.392)	0.941 (0.923)
IsFemale	2.889** (1.354)	2.332 (1.723)	2.637** (1.217)	3.497* (2.111)	2.181 (1.476)	0.151 (0.403)
Age	0.265* (0.149)	0.412** (0.189)	0.203 (0.129)	-0.030 (0.200)	0.304* (0.166)	0.108*** (0.037)
HasUnivDegree	-2.758* (1.452)	-3.652** (1.835)	-1.994 (1.304)	0.070 (2.348)	-3.188** (1.592)	-0.464 (0.394)
IsAustrian	-2.785** (1.394)	-2.897 (1.821)	-2.201* (1.227)	-2.766 (2.043)	-1.884 (1.538)	-1.013** (0.456)
Constant	85.420*** (4.121)	84.626*** (5.460)	69.623*** (3.637)	87.228*** (6.070)	60.900*** (4.675)	23.726*** (1.379)
Observations	14,890	9,890	14,890	5,000	9,890	9,890
R-squared	0.181	0.183	0.138	0.192	0.118	0.243
N Participants	1,489	989	1,489	500	989	989

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

TABLE A.9: OLS REGRESSIONS OF TIME SPENT ON INITIAL ESTIMATE

	(1)	(2)	(3)	(4)
Financial Incentives	3.437* (2.023)		3.437* (2.023)	
Algorithmic Advice	-5.774*** (1.777)	-5.771*** (1.777)		
Financial Incentives \times Algorithmic Advice	1.643 (2.433)			
PerformancePay		3.229 (2.442)		3.229 (2.442)
Tournament		3.665 (2.396)		3.665 (2.396)
PerformancePay \times Algorithmic Advice		2.366 (2.971)		
Tournament \times Algorithmic Advice		0.919 (2.922)		
AIAdvice			-6.026*** (1.951)	-6.023*** (1.951)
HumanAIAdvice			-5.547*** (2.030)	-5.544*** (2.030)
Financial Incentives \times AIAdvice			2.574 (2.716)	
Financial Incentives \times HumanAIAdvice			0.646 (2.850)	
PerformancePay \times AIAdvice				3.351 (3.401)
Tournament \times AIAdvice				1.772 (3.208)
PerformancePay \times HumanAIAdvice				1.263 (3.431)
Tournament \times HumanAIAdvice				0.046 (3.522)
Constant	77.261*** (2.368)	77.311*** (2.366)	77.265*** (2.370)	77.316*** (2.366)
Observations	14,890	14,890	14,890	14,890
R-squared	0.139	0.139	0.139	0.139
N Participants	1,489	1,489	1,489	1,489
<i>Post-hoc F tests p-values</i>				
FinInc + FinInc \times Alg. Advice = 0	0.0002			
PerfPay + PerfPay \times Alg. Advice = 0		0.0010		
Tourn + Tourn \times Alg. Advice = 0		0.0061		
FinInc + FinInc \times AIAdvice = 0			0.0009	
FinInc + FinInc \times HumanAIAdvice = 0			0.0419	
PerfPay + PerfPay \times AIAdvice = 0				0.0057
PerfPay + PerfPay \times HumanAIAdvice = 0				0.0627
Tourn + Tourn \times AIAdvice = 0				0.0109
Tourn + Tourn \times HumanAIAdvice = 0				0.1499

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

TABLE A.10: OLS REGRESSIONS OF ESTIMATION ERROR IN ALGORITHMIC
ADVICE TREATMENTS, INCLUDING WEIGHT OF ADVICE AS AN INDEPENDENT
VARIABLE

	(1)	(2)	(3)	(4)	(5)
PerformancePay	-0.323 (0.513)		-0.344 (0.510)	-0.100 (0.648)	-0.047 (0.648)
Tournament	-0.781 (0.509)		-0.795 (0.506)	-0.291 (0.683)	-0.337 (0.681)
HumanAIAdvice		-0.250 (0.413)	-0.275 (0.409)	0.197 (0.748)	0.161 (0.754)
HumanAIAdvice \times PerfPay				-0.458 (1.008)	-0.469 (1.005)
HumanAIAdvice \times Tourn				-1.000 (1.014)	-0.892 (1.018)
Weight of Advice	-3.337*** (0.481)	-3.370*** (0.486)	-3.324*** (0.481)	-3.339*** (0.486)	-3.321*** (0.487)
IsFemale = 1					0.511 (0.443)
Age					-0.008 (0.046)
HasUnivDegree = 1					0.261 (0.439)
IsAustrian = 1					-0.172 (0.484)
Constant	15.295*** (0.978)	15.038*** (0.953)	15.432*** (0.984)	15.190*** (1.018)	15.096*** (1.560)
Observations	9,711	9,711	9,711	9,711	9,711
R-squared	0.204	0.204	0.204	0.204	0.205
N Participants	989	989	989	989	989

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

TABLE A.11: OLS REGRESSIONS OF QUESTIONNAIRE ANSWERS (TABLE 6)
WITH TREATMENT INTERACTION TERMS

	(1) Task enjoyment	(2) Algorithm is credible	(3) Use of algorithm	(4) Use of context info
PerformancePay	-0.579*** (0.218)	0.201 (0.209)	7.754*** (2.652)	0.528 (1.576)
Tournament	-0.881*** (0.234)	0.400* (0.204)	8.170*** (2.450)	-0.163 (1.625)
AIAdvice	0.236 (0.205)			
AIAdvice \times PerfPay	0.252 (0.298)			
AIAdvice \times Tourn	1.044*** (0.300)			
HumanAIAdvice	0.557*** (0.187)	0.284 (0.202)	3.630 (2.499)	-0.066 (1.611)
HumanAIAdvice \times PerfPay	0.156 (0.300)	0.022 (0.279)	0.168 (3.639)	-2.402 (2.435)
HumanAIAdvice \times Tourn	0.457 (0.298)	-0.221 (0.274)	-2.384 (3.488)	1.289 (2.357)
Constant	7.700*** (0.180)	5.988*** (0.217)	46.681*** (2.497)	82.166*** (1.541)
Observations	1,489	989	989	989
R-squared	0.059	0.010	0.030	0.003

Notes: Robust standard errors in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for subject pool fixed effects.

B EXPLORING THE RELATIONSHIP BETWEEN WEIGHT OF ADVICE AND FINAL ESTIMATE ERROR

In this appendix we aim to explore why a significant increase in weight of advice due to financial incentives did not translate into a statistically significant reduction in the final estimation error, despite higher advice utilization being beneficial to performance (see also the regressions reported in Table A.10). Our suspicion is that since the final estimation error depends on several factors such as initial estimates and weight of advice, heterogeneity in these may be responsible for the (mostly) non-significance of our results on the effects of financial incentives on final estimation errors.

Our regression model 1 below estimates treatment effects on initial estimates, to verify whether incentives (and thus increased time spent on the task) have negative or positive effects on the error of participants' own initial estimate. As a second step, our approach is to separately remove heterogeneity in either initial estimates or in weight of advice, and to explore how our analysis results respond. Namely, in models 2 and 3 below we reduce heterogeneity in weight of advice (by either assuming that all participants use the same average weight of advice in their treatment, or that they employ the estimated weight of advice, respectively) while in model 4 we reduce the heterogeneity in initial estimates (by assuming all participants use the same average initial estimate for that apartment).

In Table B.1 we report results from these four regressions using data from the advice conditions. All models use treatment indicators and their interactions as independents. In order to also single out the effect of *performance pay* and *tournament incentives* within the *Human AI advice* conditions, we also report results from post-estimation f-tests on whether the joint effects $PerformancePay + HumanAIDvice \times PerformancePay$ and $Tournament + HumanAIDvice \times Tournament$ are estimated to be significantly different from zero.

Model 1 regresses the error in the initial estimate (before receiving advice) on treatment indicators. The results show that there is no evidence that financial incentives would affect the error in initial estimates. The effects for *tournament incentives* are always statistically not significant, the effect of *performance pay* is not significant with regular AI advice and even statistically significantly negative in the *Human AI* condition. This is inconsistent with a potential mechanism whereby financial incentives at the same time increase error in initial estimates (before advice) and weight of advice, such that both would cancel out each other.

The other three models use simulated final estimates as dependent variables. For the “Simulated Final Estimate 1” (regressed in model 2), we take the individual estimates of participants, but assume that all participants in a treatment condition use a weight of advice equal to the average weight of advice in their treatment condition. That is, we take out any within-treatment heterogeneity in weight of advice. We find that under this simulation, *performance pay* re-

duces error in both advice conditions while the effects for *tournament pay* are negative but not significant.

The “Simulated Final Estimate 2” (regressed in model 3) reduces noise in weight of advice in a different way. Instead of using the “raw” weight of advice calculated from comparing the individual initial and final estimate, we use the estimated weight of advice for this observation (based on model 4 presented in Table 2) in order to augment the initial estimate and arrive at the final estimate. The results are almost the same as with the first simulation: *performance pay* reduces participants’ estimate error but *tournament pay* does not.

Finally, the “Simulated Final Estimate 3” (regressed in model 4) removes a different kind of heterogeneity: the one in initial estimates. When calculating the final estimate for each participant and apartment, it uses the average initial estimate across all participants for this treatment condition and apartment, and augments it with the given advice using the participant’s individual weight of advice for this task. Under this simulation of the final estimate, *performance pay* significantly reduces error in the *AI advice* condition but has no significant effect in the *human-AI condition*, while it is the other way around for *tournament pay*: it reduces error when the AI is human-augmented but not when it is not.

The three simulation models show that heterogeneity both in terms of weight of advice and initial estimates may be responsible for the statistical null result of financial incentives on estimation error, since removing these types of heterogeneity leads to significant effects (albeit not consistently).

TABLE B.1: OLS REGRESSIONS EXPLORING EFFECTS OF WEIGHT OF ADVICE ON ERROR IN FINAL ESTIMATE ERROR

	Error Initial Estimate (1)	Error Simulated Final Estimate 1 (2)	Error Simulated Final Estimate 2 (3)	Error Simulated Final Estimate 3 (4)
PerformancePay	-0.595 (1.119)	-1.459** (0.624)	-1.335** (0.585)	-0.335** (0.138)
Tournament	0.662 (1.275)	-0.546 (0.728)	-0.394 (0.692)	-0.056 (0.142)
HumanAIAdvice	1.721 (1.345)	0.060 (0.764)	0.014 (0.705)	-0.274* (0.141)
HumanAIAdvice \times PerfPay	-2.163 (1.714)	-0.265 (0.935)	-0.241 (0.872)	0.536*** (0.199)
HumanAIAdvice \times Tourn	-2.772 (2.003)	-0.767 (1.108)	-0.856 (1.032)	-0.631*** (0.204)
Constant	33.634*** (2.555)	19.714*** (1.423)	17.600*** (1.268)	3.457*** (0.197)
Observations	9,890	9,890	9,890	9,711
R-squared	0.103	0.166	0.176	0.792
N Participants	989	989	989	989

Notes: Robust standard errors clustered by participant. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. All regressions control for apartment, order, and subject pool fixed effects.

C ADDITIONAL SURVEY STUDY ON CREDIBILITY PERCEPTIONS OF DIFFERENT FORMS OF HUMAN-AI COLLABORATIVE ADVICE

C.1 Study Design

To provide additional insights into how participants respond to the human-AI framing of the algorithmic advice, we conducted a follow-up study. Specifically, in an experimental vignette study we examine whether participants’ perceived credibility of the algorithmic advice is influenced by how the human-centered AI advice is generated. We contrast the description of the human-centered AI algorithm from our main experiment (our baseline) with five additional treatments. Each treatment describes a different mechanism of how the human experts interacted with the AI algorithm to generate the eventual human-centered AI advice.

In order to be able to freely confront participants with different mechanisms of how the human experts interacted with the AI algorithm to generate the eventual human-centered AI advice, while at the same time avoiding any deception, we employ a vignette scenario study. We first familiarized our survey participants with the Airbnb price estimation task by presenting them with an apartment listing (which as in the main experiment includes a photo, a description, the average review scores, and a map with the location) and asking them to estimate the price per night of this apartment. Every participant saw the same apartment, and the estimate was not incentivized. The main purpose of this exercise was to have participants immerse themselves into the task in order to get a feeling for the challenges encountered.

Then, we presented participants with a vignette scenario. They were asked to imagine that they no longer have to do the task on their own, but that they receive advice from a human-centered AI algorithm. We provide participants with the following description of the human-centered AI advice, which is identical to our main experiment:

Human-centered AI considers not just numbers but also human experience and advice. In fact, the human-centered AI advice in our study consists of two elements: (1) a random forest model trained on a large data set and (2) the expertise of 5 individuals who are familiar with the real-estate sector in Vienna.

(1) The random forest model is trained on a real dataset of approximately 12,000 Airbnbs in Vienna. In particular, the random forest model generates estimates of Airbnb listing prices based on an ensemble learning method for regression that operates by constructing a multitude of decision trees and returns the average estimate of the individual trees. The model takes into account the following input variables: room type (apartment, private room), number of bedrooms, number of beds, number of accommodated guests, district of Vienna, number of reviews, average review rating, and whether host is a superhost or not.

(2) *In addition to this, the price estimate of the human-centered AI incorporates advice from 5 experts. The 5 experts have substantial experience in the pricing of Airbnb apartments and are familiar with the housing and accommodation sector in Vienna.*

Our baseline scenario uses the same human-centered advice description as in our main experiment, such we do not provide any other information to participants. In five additional treatments we add another sentence to detail how the price prediction by the human experts is combined with the price prediction by the AI algorithm. The design of our new treatments is based on the hybrid intelligence concepts developed by Dellermann et al. (2019). Between-participants, we randomly assign the following additional explanations as treatment conditions:

- **Treatment 0 (baseline):** No additional text.
- **Treatment 1 (50/50 weighting):** *The eventual advice of the human-centered AI algorithm is the weighted average of the prediction of the random forest model (50% weight) and the average prediction of the 5 human experts (50% weight).*
- **Treatment 2 (80/20 weighting):** *The eventual advice of the human-centered AI algorithm is the weighted average of the prediction of the random forest model (80% weight) and the average prediction of the 5 human experts (20% weight).*
- **Treatment 3 (Human adjusts AI):** *The eventual advice of the human-centered AI algorithm is the result of a process where the random forest model first made an initial prediction which was provided to the 5 human experts, who then made the final prediction.*
- **Treatment 4 (AI adjusts human):** *The eventual advice of the human-centered AI algorithm is the result of a process where the 5 human experts first made their predictions which were then provided to the algorithm, which made the final prediction.*
- **Treatment 5 (Human-AI collaboration):** *The eventual advice of the human-centered AI algorithm is the result of a process where the 5 human experts and the algorithm interacted with each other to make the final prediction.*

As our main dependent variable, we ask participants to what extent they would find the advice of such a human-centered AI algorithm to be a credible source for estimating the price of an Airbnb apartment. This question on source credibility, which is based on Chen et al. (2022), is identical to an item in our post-experiment questionnaire of the main experiment. Participants respond on a scale from 0 (to no extent) to 10 (to a very large extent). After the

measurement of our dependent variable, on the next survey page we employed an ex-post recall check by asking participants to remember the exact description of the humanAI algorithm. The survey ended with a few brief questions on demographics.

We randomized the presentation format since it may affect how rather complex information is processed by participants. Half of the participants received all the information on the human-centered AI algorithm on one single page, while the remaining half received the information sequentially on multiple screens. We control for this information provision format in our subsequent analyses.

We recruited participants from a large public university in Austria via a university-wide survey mailing list. In total, we received 1601 responses, from which we excluded 72 participants for suspected double participation from the same IP address, and 13 participants who did not provide their contact details upon the study’s completion. Thus, our final dataset includes observations from 1,516 participants. In terms of demographics, the average participant age is 22.3 years and 52% identify as female. About one third are masters degree students, and the remaining two thirds are undergraduate students in business, economics, or law. 54% are Austrian nationals, while more than 70 different nationalities are present in the remaining half. As a “thank you” for participating in our survey study, participants could enter a lottery, with five winners each receiving EUR 100.

C.2 Results

In Table C.1, we present descriptive statistics on participants’ perceived credibility in the human-centered AI advice. Columns 1-3 present descriptives for the full sample, while columns 4-6 report aggregates only for those participants who passed the ex-post recall check. Due to the complex nature and lengthy description of the human-centered AI algorithms, a rather large fraction of participants failed the ex-post recall check in our study. However, we do not observe any significant differences in our study results between the two sample groups.

Table C.2 reports results from OLS regressions of the perceived credibility of the human-centered AI advice on our five treatments conditions. Columns 1-3 refer to the whole sample, columns 4-6 only consider participants who passed the ex-post recall check. Columns 1 and 4 are simple OLS models, columns 2 and 5 control for the presentation format (sequential vs. simultaneous), and columns 3 and 6 controls for participant demographics. Across all specifications, we do not observe any statistically significant differences between our baseline treatment and the five additional treatments. Also across the five treatments, we do not observe a clear pattern that some human-AI collaborations would be rated higher or lower on credibility than others. In fact, across all treatments, participants rate the advice credibility with the same value of around 6 out of 10.

TABLE C.1: AVERAGE CREDIBILITY ACROSS TREATMENTS

	Full sample			Passed recall check		
	N	Avg.	(StdDev)	N	Avg.	(StdDev)
Treatment 0: Baseline	259	6.205	(1.765)	259	6.205	(1.765)
Treatment 1: 50/50 weighting	247	6.170	(1.690)	185	6.249	(1.682)
Treatment 2: 80/20 weighting	255	6.157	(1.716)	210	6.181	(1.738)
Treatment 3: Human adjusts AI	262	6.137	(1.750)	132	6.136	(1.716)
Treatment 4: AI adjusts human	261	6.061	(1.729)	94	5.989	(1.569)
Treatment 5: Human-AI coll.	232	6.091	(1.806)	87	6.023	(1.874)
All	1,516	6.137	(1.740)	967	6.161	(1.726)

This suggests that the way in which the human expert predictions and the AI predictions are combined does not meaningfully affect the perceived credibility of the human-centered AI advice. Thus, we conclude that our inferences from the main experiment are likely generalizable to many different settings in practice, regardless of how specifically human and machine expertise is combined to generate a human-centered AI advice.

TABLE C.2: AVERAGE CREDIBILITY ACROSS TREATMENTS

	Full sample			Passed recall check		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment = 1	-0.035 (0.155)	-0.034 (0.155)	-0.030 (0.154)	0.044 (0.166)	0.039 (0.167)	0.045 (0.166)
Treatment = 2	-0.048 (0.154)	-0.054 (0.154)	-0.048 (0.153)	-0.024 (0.161)	-0.031 (0.161)	-0.033 (0.160)
Treatment = 3	-0.067 (0.153)	-0.066 (0.153)	-0.068 (0.152)	-0.068 (0.185)	-0.075 (0.185)	-0.087 (0.185)
Treatment = 4	-0.143 (0.153)	-0.140 (0.153)	-0.158 (0.152)	-0.215 (0.208)	-0.213 (0.208)	-0.232 (0.208)
Treatment = 5	-0.114 (0.157)	-0.114 (0.157)	-0.115 (0.157)	-0.182 (0.214)	-0.177 (0.214)	-0.184 (0.214)
Sequential		0.098 (0.090)	0.089 (0.089)		0.089 (0.112)	0.072 (0.112)
Age			-0.017 (0.014)			-0.006 (0.018)
Austrian			-0.058 (0.091)			-0.032 (0.113)
Female			-0.274*** (0.090)			-0.254** (0.111)
MA degree			0.257*** (0.099)			0.251** (0.126)
Constant	6.205*** (0.108)	6.155*** (0.117)	6.364*** (0.276)	6.205*** (0.107)	6.160*** (0.121)	6.104*** (0.342)
Observations	1,516	1,516	1,516	967	967	967
R-squared	0.001	0.002	0.013	0.002	0.003	0.015

Notes: The dependent is the credibility rating [1-10] of the humanAI algorithm. Columns 1-3 report results from the full sample, columns 4-6 only include participants who passed the recall check. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

D THE ALGORITHM UNDERLYING THE ALGORITHMIC ADVICE IN THE EXPERIMENT

The algorithm used for the experimental stimuli (the algorithmic advice) is based on a random forest model which is trained on a real dataset obtained from www.insideairbnb.com. The raw data set of Vienna listings was originally scraped on 09.06.2021 and contained 11,567 observations. The model generates estimates of Airbnb listing prices based on a random forest model – an ensemble learning method for regression that operates by constructing a multitude of decision trees and returns the average estimate of the individual trees. We include the full code of the model in the replication package.

D.1 Data cleaning and variables

The dataset contains the following information: numerical ID (`id`), weblink to actual Airbnb listing (`listing_url`), listing title (`name`), description of Apartment (`description`), title picture of the apartment (`picture_url`), yes/no indication if host is an “Airbnb Superhost” (`host_is_superhost`), yes/no indication if host’s identity was verified (`host_identity_verified`), the district in Vienna where the apartment is located (`neighbourhood_cleansed`), an indicator whether an entire home/apartment or just a room was offered (`room_type`), the number of guests than can be accommodated (`accommodates`), the number of bathrooms (`bathrooms_text`), the number of bedrooms (`bedrooms`), the number of beds (`beds`), the apartment base price per night (`price`), the total number of reviews (`number_of_reviews`), the number of reviews last three months (`number_of_reviews_ltm`), the number of reviews last 30 days (`number_of_reviews_l30d`), the date of the first review (`first_review`), the date of the last review (`last_review`), the overall review score (`review_scores_rating`), and the review scores on sub-categories accuracy, cleanliness, check-in, communication, location, and value (`review_scores_XX`).

Before the dataset is used in the model, it is cleaned and reduced. In particular, observations of hotel rooms or shared rooms, where superhost status was not known, where prices were outliers (prices smaller than 30 USD and above 300 USD), or where hosts were very inexperienced (less than 5 reviews) were dropped. The final dataset for training the algorithm consists of 5,426 observations.

The algorithm takes into account the following input variables: room type (apartment, private room), number of bedrooms, number of beds, number of accommodated guests, district of Vienna, number of reviews, average review rating, and whether host is a superhost or not. The algorithm deploys an 80/20 stratified sampling split, so 80% of the data was used as a training set and 20% as a test set.

D.2 Model specifications and selection

Our algorithm compares decision-tree models and random forest models. It is implemented in *R*. First, we perform grid-based hyperparameter search for a decision tree model. We use a grid of potential values for the hyperparameter(s) that we want to try – we specify potential values for the hyperparameters, and the *tune_grid()* function builds separate models using these values. We compare how well the individual values work on the cross-validated data set and select the "best" set of hyperparameters to predict the separate test data set. To include grid search in our pipeline, we specify three things: a) in the *decision_tree()* function we specify which hyperparameters we want to tune, b) we define a grid of hyperparameter values and provide it to the *fit* function, c) we use the function *tune_grid()* to tell *tidymodels* that we want to tune the previously defined hyperparameters. We then select the hyperparameters based on the best Root Mean Squared Errors, and predict the test set.

Second, we use random search as an efficient alternative to grid search in decision tree modeling. In random search, we switch our grid of hyperparameter values to *grid_random*, and *R* creates random values for the hyperparameters to try. We specify a random grid using 10 different combinations of values. We then compare the results of the random search with the results of the grid search. We select the best hyperparameters according to the metric Root Mean Squared Error and predict the test set. Then we compare our decision-tree models based on grid and random search.

Third, we specify a random forest model. Random Forests are a type of bagging approach where multiple, independent decision trees are built on separate, independent, bootstrapped data sets. Random forests improve upon standard bagging approaches by de-correlating the individual trees. If we have a set of very strong predictors in the data set, these predictors will be used over and over by the separate trees. To prevent this, random Forests randomly select a subset of predictors at each step that are considered for determining the next split in the tree. This guarantees that "less important" predictors are chosen sometimes. In *R*, this hyperparameter is called *mtry*. We try different values for this parameter and select the one that works best for our dataset. Additional parameters that we tune are *trees* (the number of trees) and *min_n* (minimum number of data points). To specify cross-validation, we use the *vfold_cv* function. We predict the test set.

Finally, we can compare the predictions of decision tree and random forest models on the test set. We a) sort the results according to a specific metric Root Mean Squared Error, b) extract the best model type, c) extract the best hyperparameters for this model type, d) re-train the final model using the best hyperparameters, and e) predict the test data. Ultimately the random forest model is chosen based on this metric.

E SCREENSHOTS OF EXPERIMENTAL INSTRUCTIONS AND DECISION SCREENS

Screenshot 1: Welcome screen, payment information, and consent form

WELCOME!

Thank you for participating. In this study, we try to better understand how people form economic judgements and make decisions.

The study will take around **15-20 minutes**. Your main task will be to submit price estimates for past Airbnb listings in Vienna. This will be followed by a questionnaire.

100 individuals out of all participants in this study will be selected for payment. If you are one of these 100 randomly selected individuals, you can earn up to **EUR 100**. (The average payment for paid participants will be around EUR 50.) The study is funded by WU Vienna.

If you are one of the 100 participants chosen for payment, we will inform you by email about your payment. In order to do so, we will ask you to provide us with your name and email address at the end of the study. The randomly selected participants will then be asked for their IBAN account number, and paid by WU via IBAN bank transfer. Once this is completed, any identifying data will be removed from the dataset, including names, email addresses, and IBAN numbers. Only the WU Financial Accounting department (Finanzbuchhaltung) will keep records of the transfers as required by law.

Irrespective of you being chosen for payment or not, all your answers and decisions in this task will be treated **confidentially**. Results of this study will only be presented in aggregated form.

Your participation is **completely voluntary** and you can stop participating at any time and for any reason or for no reason at all. (However, you will not be eligible for payment if you do not complete the study.)

Contact: If you have any questions or comments about this study, please feel free to contact Georg Lintner at georg.lintner@wu.ac.at.

Please indicate below that you agree to participate in our study:

- ☐ I agree to participate.
- ☐ I do not agree to participate.

Screenshot 2: Filtering question

In the next 15-20 minutes, I am able to fully concentrate on this study.

- ☐ Yes.
- ☐ No.

Screenshot 3: Ground rules of the study

Before you start, please consider the following ground rules:

- **Participating multiple times in the study is strictly forbidden!** Doing so will lead to an immediate disqualification and you will not receive any payment!
- Please **do not communicate** with other individuals while participating in the study!
- **We promise not to misrepresent facts or deceive you in any way!** All the information provided in the study is 100% true and your performance as well as your payments will be determined exactly as described.

Screenshot 4: Basic task description

Your task

Your task in this experiment will be to **estimate the price per night of 10 different Airbnbs in Vienna**. The price per night in our task is defined as the **base price excluding any additional fees** (i.e. excl. Airbnb service fee and/or cleaning fees).

"Airbnbs" are apartments or private rooms rented for short-term tourism or business stays via the online platform Airbnb.com.

The 10 different Airbnb listings were **real listings in Vienna in June 2021**, but they are **currently not active** anymore. Hence, looking for the prices online will not be effective, and we ask you not to do so for time reasons. We are interested in **your** estimates.

Screenshot 5: Detailed task description and comprehension questions

Airbnb price estimation

To make your estimate, we will provide you with the actual Airbnb listing in Vienna, which includes a photo, a brief description of the apartment, the approximate location, the average rating by previous guests, and some information on the host. You will see an example for a typical Airbnb listing in a moment.

The label "entire apartment" refers to renting the whole apartment - no facilities and rooms are shared with the host or other guests. The "private room" label refers to renting a room in a shared apartment. Kitchen and bathroom are usually shared with the host or other guests.

Please note that there might be some **attention checks** included in the task. An attention check will look similar to the 10 apartment listings, but will ask you in the text to submit a particular number. In case you fail an attention check, you will not receive any payment.

For our research purposes we rely on you that you try to make as accurate price estimates as possible.

Your task is to estimate the price per night of 10 Airbnb listings in Vienna.

- ☐ True.
- ☐ False.
- ☐ I don't know.

Your task is to make as accurate price estimates as possible.

- ☐ True.
- ☐ False.
- ☐ I don't know.

Screenshot 6: Example visualization of an Airbnb listing

This is how the information you will receive for each Airbnb will look like:

Listing cover photo

About this space
[description of the listing; max. 600 characters]

[Listing title]

[Room type] in [district]

of guests - # of bedrooms - # of beds - # of baths

No superhost

Host identity not verified

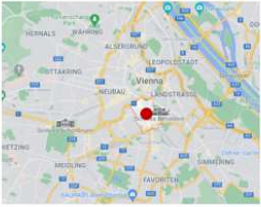
Superhost

Host identity verified

Avg. review scores (5.0 = maximum)

★ X.XX (# of reviews)

Accuracy	X.X	Communication	X.X
Cleanliness	X.X	Location	X.X
Check-in	X.X	Value	X.X



Airbnb is within the red-framed area

Screenshot 7: Detailed instructions on the AI Algorithm (only AI treatments)

Decision Support by an Artificial Intelligence (AI) Algorithm

To support you in making your estimates, we will also provide you with an advice that was generated by an artificial intelligence algorithm.

The algorithm itself is based on a **random forest model** which is trained on a real dataset of approximately 12,000 Airbnbs in Vienna (as of June 2021). In particular, the random forest model generates estimates of Airbnb listing prices based on an ensemble learning method for regression that operates by constructing a multitude of decision trees and returns the average estimate of the individual trees. The algorithm takes into account the following input variables: room type (apartment, private room), number of bed rooms, number of beds, number of accommodated guests, district of Vienna, number of reviews, average review rating, and whether host is a superhost or not.

However, please be advised that **the algorithm is not perfect. Its estimates can be above or below the actual listing price.** For some Airbnbs, the algorithm produces relatively accurate estimates, for some others slightly deviating estimates, and for some Airbnbs its estimates can be far off the actual listing price. Based on previous analysis, on average the **algorithm is about 30% off the true price.**

For each of the 10 Airbnb listings, you will first be asked to make an **initial price estimate without having the algorithmic advice.** Only then you will receive the advice from the AI algorithm, and you are asked to make a **second price estimate.** The second estimate can be equal or different to your first estimate. So for each Airbnb listing you will be asked to make two estimates.

The AI algorithm is trained on a real dataset of Airbnb listings in Vienna.

- ☐ True.
- ☐ False.
- ☐ I don't know.

The AI algorithm is based on a

- ☐ Random forest algorithm.
- ☐ Support vector machine.
- ☐ Simple linear regression.

The average estimation error of the AI algorithm is

- ☐ 0%
- ☐ 30%
- ☐ 50%

Screenshot 8: Detailed instructions on the Human-framed AI Algorithm

Decision Support by a Human-Centered Artificial Intelligence (AI) Algorithm

To support you in making your estimates, we will also provide you with an **advice that was generated by a human-centered artificial intelligence algorithm**. Human-centered AI takes into account not just numbers but also human experience and advice.

The algorithm itself is based on a **random forest model** which is trained on a real dataset of approximately 12,000 Airbnbs in Vienna (as of June 2021). In particular, the random forest model generates estimates of Airbnb listing prices based on an ensemble learning method for regression that operates by constructing a multitude of decision trees and returns the average estimate of the individual trees. The algorithm takes into account the following input variables: room type (apartment, private room), number of bed rooms, number of beds, number of accommodated guests, district of Vienna, number of reviews, average review rating, and whether host is a superhost or not.

In addition to this, the price estimate incorporates expert advice from 5 individuals. The **5 experts have substantial experience** in the pricing of Airbnb apartments and are familiar with the housing and accommodation sector in Vienna.

However, please be advised that **the human-centered algorithm is not perfect. Its estimates can be above or below the actual listing price**. For some Airbnbs, the algorithm produces relatively accurate estimates, for some others slightly deviating estimates, and for some Airbnbs its estimates can be far off the actual listing price. Based on previous analysis, on average the **algorithm is about 30% off the true price**.

For each of the 10 Airbnb listings, you will first be asked to make an **initial price estimate without having the algorithmic advice**. Only then you will receive the advice from the AI algorithm, and you are asked to make a **second price estimate**. The second estimate can be equal or different to your first estimate. So for each Airbnb listing you will be asked to make two estimates.

The estimate by the decision support system is generated by a human-centered AI algorithm.

- ☐ True.
- ☐ False.
- ☐ I don't know.

The human-centered AI algorithm is trained on

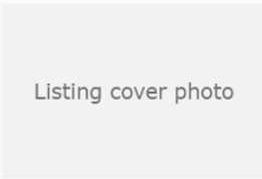
- ☐ A dataset of 12,000 Airbnb listings.
- ☐ Advice of 5 experts.
- ☐ Both: a dataset of 12,000 Airbnb listings and the advice of 5 experts.

The average estimation error of the AI is

- ☐ 0%
- ☐ 30%
- ☐ 50%

Screenshot 9: Example of Airbnb listing with algorithmic advice



After your first estimate, you will receive the same listing. Now also with the price estimate by the algorithm.





[Listing title]

[Room type] in [district]


of guests - # of bedrooms - # of beds - # of baths

 No superhost
  Host identity not verified

 Superhost
  Host identity verified


About this space

[description of the listing; max. 600 characters]

 X.XX (# of reviews)

Avg. review scores (5.0 = maximum)

Accuracy	X.X	Communication	X.X
Cleanliness	X.X	Location	X.X
Check-in	X.X	Value	X.X



Airbnb is within the red-framed area

Price estimate by algorithm: XX €

[information about how the algorithm works]

Screenshot 10: Compensation contract details – fixed payment

Your compensation:

100 individuals out of all participants in this study will be randomly selected for payment.

If you are one of these randomly selected participants, you will receive a **fixed payment of EUR 50 for providing us with your estimates.**

How will you be paid for doing the price estimation tasks?

- ☐ If I am one of the 100 randomly selected participants, I will receive no payment.
- ☐ If I am one of the 100 randomly selected participants, I will receive a payment of EUR 50.
- ☐ If I am one of the 100 randomly selected participants, I will receive a payment based on my performance in estimating Airbnb listing prices.

Screenshot 11: Compensation contract details – tournament incentives (as shown to treatment groups without algorithmic advice)

Your Compensation:

100 individuals out of all participants in this study will be randomly selected for payment.

If you are one of these randomly selected participants, you can earn **EUR 100 conditional on your performance compared to another participant**. We will randomly select one of your 10 price estimates. You will be randomly matched with another participant who is also selected for payment, and your price estimate is compared to the other participant's price estimate. **If your estimate is closer to the true price, you receive EUR 100** and the other participant receives EUR 0. Vice versa, if the other participant's estimate is closer to the true price, then the other participant receives EUR 100 and you receive EUR 0. If you submitted exactly the same price estimate, then it will be randomly determined who receives EUR 100 and who receives EUR 0.

How will you be paid for doing the price estimation tasks?

- ☐ If I am one of the 100 randomly selected participants, I will receive no payment.
- ☐ If I am one of the 100 randomly selected participants, I will receive a payment of EUR 50.
- ☐ If I am one of the 100 randomly selected participants, I will receive a payment of EUR 100 if my price estimate is more accurate than the price estimate of another randomly matched participant.

Screenshot 12: Compensation contract details – performance-based incentives (as shown to treatment groups with algorithmic advice)

Your Compensation:

100 individuals out of all participants in this study will be randomly selected for payment.

If you are one of these randomly selected participants, **you can earn EUR 100 for your price estimation**. The computer will randomly select one of your 20 price estimates (you will make 2 estimates for each listing), and your likelihood to receive a payment of **EUR 100** will depend on how close your price estimate is to the true price. **The better your price estimate, the higher are your chances to receive the EUR 100**, such that your optimal action is to submit your best estimate of the price.

In particular, we will calculate a score between 0 and 100 as follows:

$$\text{Score} = \max (100 - 0.2 \times (\text{Estimate} - \text{TruePrice})^2 , 0)$$

Based on this function, you can approximately expect the following scores contingent on how much your estimate deviates from the true price:

Estimate deviation	Score (=likelihood to receive EUR 100)
EUR 0	100
EUR 5	95
EUR 10	80
EUR 15	55
EUR 20	20
EUR 22	12
EUR 24	3
>=EUR 25	0


The computer will randomly draw a number between 0 and 100 (with each number being equally likely). If your score is higher than or equal to that random number, you will receive the EUR 100. If your score is lower than the random number, you will receive EUR 0 (nothing). Thus, your score directly represents your likelihood (in percent) to receive EUR 100, and the higher your score, the higher are your chances to receive the EUR 100.

How will you be paid for doing the price estimation tasks?

☐ If I am one of the 100 randomly selected participants, I will receive no payment.
 ☐ If I am one of the 100 randomly selected participants, I will receive a payment of EUR 50.
 ☐ If I am one of the 100 randomly selected participants, I will receive a payment based on my performance in estimating Airbnb listing prices.

Screenshot 13: Attention check (same for all treatment groups)

Attention check



Attention check apartment in Vienna

Attention check: Please submit "1000" as price

0 guests - 5 bedrooms - 1 bed - 2 baths

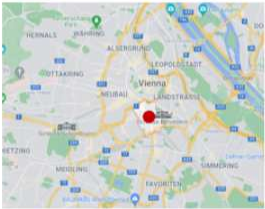
No superhost Host identity not verified

About this space

If you are reading this, you are still paying adequate attention. To indicate this, please submit "1000" as your guess for this listing. This listing will not count towards your performance. Thank you for your ongoing attention.

4.55 (100 reviews)

Accuracy	5.0	Communication	2.0
Cleanliness	3.0	Location	3.7
Check-in	4.5	Value	5.0




Airbnb is within the red-framed area

Your estimated base price (as of Jun21; excl. Airbnb service fee and/or cleaning fees) for this listing in EUR per night:

Screenshot 14: Example Airbnb listing – without algorithmic advice (treatment groups)

Listing information



Lovely designed Apartment - 5 Minutes from Center

Entire apartment in Alsergrund

2 guests - 1 bedroom - 1 bed - 1 bath

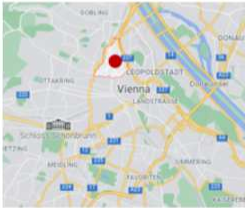
Superhost Host identity verified

About this space

Completely new and lovely renovated CITY Apartment (40sqm) in the 9th district very close to the inner center - only 10 min. to walk and 5 Minutes per tram. So its a ideal apartment for sightseeing. Nevertheless the apartment is in a quiet side street, and lots of good and charming restaurants nearby. 1 Livingroom with dining place and working place 1 Sleeping Room with Queen-Size Bed 1 Kitchen with Coffee maschine and dishwasher 1 Bathroom with walk in shower 1 seperate Toilette WLAN The Apartment is completely new renovated and furnished. ...

4.86 (88 reviews)

Accuracy	5.0	Communication	4.9
Cleanliness	4.9	Location	4.7
Check-in	5.0	Value	4.8



Airbnb is within the red-framed area

Your estimated base price (excl. additional fees such as cleaning fees) for this listing in EUR per night:

Screenshot 15: Example Airbnb listing – with AI advice

Listing information



Lovely designed Apartment - 5 Minutes from Center

Entire apartment in Alsergrund

2 guests - 1 bedroom - 1 bed - 1 bath



Superhost



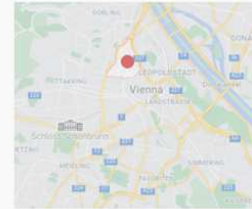
Host identity verified

About this space

Completely new and lovely renovated CITY Apartment (40sqm) in the 9th district very close to the inner center - only 10 min. to walk and 5 Minutes per tram. So its a ideal apartment for sightseeing. Nevertheless the apartment is in a quiet side street, and lots of good and charming restaurants nearby. 1 Livingroom with dining place and working place 1 Sleeping Room with Queen-Size Bed 1 Kitchen with Coffeemaschine and dishwasher 1 Bathroom with walk in shower 1 separate Toilette WLAN The Apartment is completely new renovated and furnished. ...

★ 4.86 (88 reviews)

Accuracy	5.0	Communication	4.9
Cleanliness	4.9	Location	4.7
Check-in	5.0	Value	4.8



● Airbnb is within the red-framed area

Price prediction by machine learning algorithm (random forest): 59 €


The AI advice was developed based on numerical input of ~12,000 listings in Vienna. The highly advanced random forest algorithm generates forecasts of Airbnb listing prices based on an ensemble learning method for regression that operates by constructing a multitude of decision trees and returns the average prediction of the individual trees.



Your estimated base price (as of Jun21; excl. Airbnb service fee and/or cleaning fees) for this listing in EUR per night:

Screenshot 16: Example Airbnb listing – with human-framed AI advice

Listing information



Lovely designed Apartment - 5 Minutes from Center

Entire apartment in Alsergrund

2 guests - 1 bedroom - 1 bed - 1 bath

Superhost


Host identity verified

About this space

Completely new and lovely renovated CITY Apartment (40sqm) in the 9th district very close to the inner center - only 10 min. to walk and 5 Minutes per tram. So its a ideal apartment for sightseeing. Nevertheless the apartment is in a quiet side street, and lots of good and charming restrauments nearby. 1 Livingroom with dining place and working place 1 Sleeping Room with Queen-Size Bed 1 Kitchen with Coffee maschine and dishwasher 1 Bathroom with walk in shower 1 seperate Toilette WLAN The Apartment is completely new renovated and furnished. ...

★ 4.86 (88 reviews)


Accuracy	5.0	Communication	4.9
Cleanliness	4.9	Location	4.7
Check-in	5.0	Value	4.8



Airbnb is within the red-framed area

Price prediction by human-centered AI system: 59 €

The AI advice was developed based on numerical input of ~12,000 listings in Vienna as well as predictions by human experts. Human-centered AI is defined by systems that are continuously improving because of human input while providing an effective experience between human and algorithm.



Your estimated base price (as of Jun21; excl. Airbnb service fee and/or cleaning fees) for this listing in EUR per night:

67

Screenshot 17: Post experiment questionnaire – demographics

In the following sections, we will ask you a few questions about yourself. Your answers will be treated confidentially and anonymously. Please be honest, this is very important for our research.

How old are you, in years?

What is your nationality?

What gender do you identify with?

- ☐ Female
- ☐ Male
- ☐ Other
- ☐ I do not want to answer

What is your highest completed education?

- ☐ Middle/high school
- ☐ Bachelor's degree
- ☐ Master's degree
- ☐ PhD degree or higher

How familiar are you with Vienna?

- ☐ I live in Vienna
- ☐ I don't live in Vienna but have been there many times.
- ☐ I don't live in Vienna but have been there a few times.
- ☐ I have never been to Vienna.

On a scale from 0 to 10, how well do you know Vienna?

0 not at all 10 very well

0 1 2 3 4 5 6 7 8 9 10

Have you used Airbnb in the past?

- ☐ I have previously booked an Airbnb in Vienna
- ☐ I have previously booked an Airbnb elsewhere but not in Vienna
- ☐ I have never booked an Airbnb


Screenshot 18: Post experiment questionnaire – Risk taking measure

Please tell us, in general, how willing or unwilling you are to take risks.

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

0 completely unwilling to take risks 10 very willing to take risks

0 1 2 3 4 5 6 7 8 9 10




Screenshot 19: Post experiment questionnaire – Task enjoyment measure

How much did you enjoy the task of estimating prices of Airbnb listings.

Please use a scale from 0 to 10, where 0 means you "Did not enjoy it at all" and 10 means "Did enjoy it a lot". You can use any number between 0 and 10 to indicate where you fall on the scale, like 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

0 did not enjoy it at all 10 did enjoy it a lot

0 1 2 3 4 5 6 7 8 9 10




Screenshot 20: Post experiment questionnaire – Overconfidence measure

Across all 10 listings, what do you think is your average deviation from the actual listing price?

0 10 20 30 40 50 60 70 80 90 100

%



Screenshot 21: Post experiment questionnaire – Information reliance measures (bottom question only shown in algorithmic advice treatments)

Across all 10 listings, how much did you rely on the contextual information (i.e., picture, title, description, map)? (0% not considering it at all, 100% fully relied on it)

0 10 20 30 40 50 60 70 80 90 100

%

Across all 10 listings, how much did you rely on the algorithmic advice? (0% not considering it at all, 100% fully relied on it)

0 10 20 30 40 50 60 70 80 90 100

%

Screenshot 22: Post experiment questionnaire – Unfamiliarity questions

Please indicate how much you agree or disagree with the following statements about **potential challenges in the task**.

I found it difficult to evaluate which areas in Vienna are good to stay in for tourists.

0 Strongly disagree 10 Strongly agree

0 1 2 3 4 5 6 7 8 9 10

I was not aware of the price level for renting apartments in Vienna.

0 Strongly disagree 10 Strongly agree

0 1 2 3 4 5 6 7 8 9 10

I found it hard to assess how an average apartment in Vienna looks like.

0 Strongly disagree 10 Strongly agree


0 1 2 3 4 5 6 7 8 9 10

Screenshot 23: Post experiment questionnaire – Source credibility measure and free text (only shown in algorithmic advice treatments)

To what extent do you find the advice by the algorithm to be a credible source for estimating the prices of Airbnb apartments?

0 To no extent 10 To a very large extent

0 1 2 3 4 5 6 7 8 9 10



When your second estimate was different to the algorithmic advice, please describe your main reason(s) for not following the algorithmic advice.

Screenshot 24: Final screen and contact details form

Thank you for participating in this study.

In order to be able to contact you in case you are one of the 100 participants who are randomly selected for payment, please provide your name and email address below. Otherwise, without your details we will not be able to inform you about your payment. We will use your e-mail address only for this study's payment purposes and delete it afterwards.

We plan to facilitate all IBAN transfers for the 100 randomly selected participants in the week before Christmas.

What is your name?

What is your email address?

Contact:
For questions or comments about this study, please contact Georg Lintner (georg.lintner@wu.ac.at).