INSEAD The Business School for the World®

Working Paper

2025/16/DSC

Does Autonomy Make People Try Less Hard? Initial Weight Inertia in Human-AI Collaboration

Qiong Xia
INSEAD, qiong.xia@insead.edu

Vivek Choudhary
Nanyang Business School NTU, vivek.choudhary@ntu.edu.sg

Rajat Sharma
Indian Institute of Management, rajats@iima.ac.in

Yash Raj Shrestha
University of Lausanne, yashraj.shrestha@unil.ch

Organizations increasingly adopt AI tools in decision-making, yet humans often exhibit AI aversion and disregard its inputs. While granting managers full autonomy in weighting AI recommendations can address this aversion, it risks limiting potential gains if not used effectively, leading to lower performance. Through an online experiment where participants are tasked with making collaborative predictions leveraging AI advice, we examine this autonomy-performance trade-off. Our results reveal that participants with full autonomy exhibit systematic *initial weight inertia* characterized by anchoring to initial AI weights and a ceiling effect where users fail to exceed initial weights put on AI advice. Participants also disregard performance feedback, failing to adjust weights even when consistently underperforming compared to AI. To mitigate this inertia while preserving autonomy, we develop a *bounded autonomy approach*, constraining weight adjustments based on collaborative performance relative to AI's predictions alone. When AI performs well, participants cannot fully disregard its recommendations, especially when their own predictions are inaccurate. This approach reduces human-AI collaborative prediction errors by 12.47%, reducing *initial weight inertia* and improving feedback responsiveness. Finally, we find that different framings (weighting AI vs. weighting participants' own predictions) yield similar performance improvements but via distinct mechanisms: weighting AI predictions increases cognitive effort, while weighting original predictions eliminates anchoring bias. This novel approach balances user autonomy with improved performance, offering a practical solution for enhancing human-AI collaboration.

Keywords: Behavioral Operations; Autonomy; Initial Weight Inertia, Human-AI Collaboration; Experiment

Electronic copy available at: https://ssrn.com/abstract=5068677

1. Introduction

In the rapidly evolving landscape of technological innovation, human-AI collaboration for prediction has gained significant attention (Ibrahim et al. 2021, Snyder et al. 2024, Balakrishnan et al. 2024). As artificial intelligence (AI) technology continues to advance, businesses across various domains, including operations, increasingly employ AI algorithms to enhance critical decision-making processes (Raisch and Krakowski 2021). While AI often generates high-quality predictions that surpass human capabilities in predictive accuracy (Dietvorst et al. 2015, Kellogg et al. 2020), humans add critical value by significantly reducing extreme errors, especially when they possess private information (Ibrahim et al. 2021, Kim and Song 2022, Balakrishnan et al. 2024, Chen et al. 2023). Despite these benefits, a paradoxical phenomenon has been observed: in practice, humans often disregard AI inputs and fail to benefit from potentially highly accurate AI predictions – a phenomenon known as algorithm aversion (Dietvorst et al. 2015).

Research has explored various approaches to mitigate algorithm aversion, with a promising direction involving human autonomy in deciding how much to rely on AI predictions. For instance, studies have found that allowing humans to freely override AI predictions effectively reduces algorithm aversion (Dietvorst et al. 2018, Fink et al. 2024). However, this approach reveals a critical trade-off: while autonomy enhances AI acceptance by allowing greater control over AI predictions, it may simultaneously undermine the accuracy of collaborative decision-making. A prime example illustrates this tension: overconfident managers might override accurate AI predictions, thereby failing to fully leverage the AI's value (Dargnies et al. 2024). Despite this evident tension between autonomy benefits and decision quality risks in human-AI collaboration, a systematic exploration of this trade-off remains unexplored in the existing literature.

In this paper, we systematically explore these trade-offs associated with granting users full autonomy in integrating AI predictions in a sequential prediction task, where humans and AI produce their predictions in sequence which is subsequently integrated (e.g., Balakrishnan et al. 2024). Specifically, we address two critical research questions: (1) how does full autonomy to use AI predictions impact weighting strategies and overall collaborative performance? (2) what strategies can mitigate the adverse effects of granting full autonomy? To address these questions, we propose a novel *bounded autonomy approach* that strategically limits user control over their reliance on AI predictions based on demonstrated performance. We posit that this approach improves prediction accuracy, encourages better weighting strategies, and facilitates learning from feedback, enabling users to enhance their ability to collaborate with AI systems and ultimately become effective predictors over time.

To answer our research questions, we design and conduct an online experiment featuring a sequential prediction task. For each prediction round, participants first made an original prediction by themselves. In the experimental groups, participants were then shown the AI's prediction and produced their revised predictions (which is collaborative performance, incorporating AI advice when available) by assigning a weight (from 0–100%) to the AI's prediction while combining it with their own original prediction. A 0% weight reflects complete reliance on their original prediction, while a 100% weight indicates complete reliance on the AI's prediction. In the control group, participants revised their predictions without any AI assistance. After each round, participants received performance feedback, including the absolute difference between their revised predictions and the actual outcomes (ground truths). We measured performance using the established metric of absolute errors (AE) (Dietvorst et al. 2018).

In our analysis of how people assign weight to AI predictions under full autonomy, our experiment reveals surprising results. *There is a downside to autonomy*: participants exhibit *initial weight inertia* in integrating AI predictions in revising their predictions, despite AI in our experiment shown to significantly improve their performance.[1] This behavior is strongly anchored by their initial weights (typically around 50%). Despite receiving performance feedback in subsequent rounds, participants continue to anchor on these initial weights rather than adjusting them based on their actual performance. Moreover, we find that most participants assign their maximum weight to AI predictions in the first round and never go beyond it, indicating a ceiling effect that further demonstrates the persistent impact of initial weight selections. This suggests that while AI assistance improves overall performance, yet participants show limited learning in two critical ways: they neither improve their original predictive accuracy over time nor learn to better estimate weights to place on AI predictions incorporating the performance feedback.

To address this side effect of full autonomy, we develop a bounded autonomy approach that dynamically adjusts the range of control users have over the weighting of AI predictions based on their performance. Under this approach, users who achieve better collaborative performance earn greater freedom to determine how much they want to rely on AI predictions versus their original predictions, while those who perform poorly face more constraints. Therefore, *better collaborative performance rewards the users with more decision-making autonomy.* This approach maintains user agency while introducing guardrails that encourage more thoughtful integration of AI predictions.

We find that the bounded autonomy approach to weighting AI predictions reduces errors by 12.47% compared to unbounded autonomy. This improvement is achieved through reduced *initial weight inertia* - manifested as lower anchoring and ceiling effects of initial weight assignments

[1] We find that prediction errors decrease by 34.36% when participants are provided with AI advice

on subsequent weighting decisions, and increased responsiveness to performance feedback. These mechanisms work together to overcome the systematic *initial weight inertia* observed in our study. To further explore the effectiveness of bounded autonomy, we test an alternative implementation with a different framing, where participants were asked to assign weights to their original predictions instead of AI predictions. While both implementations achieve similar accuracy improvements (12.04% for bounded autonomy in weighting original predictions), they operate through distinct mechanisms. Specifically, when participants are asked to assign weights to their original predictions, this effectively eliminated the anchoring effect. In contrast, when participants are asked to assign weights to AI predictions, their cognitive effort increases during the weighting decisions.

Our research contributes to the understanding of human-AI collaboration in operations management by identifying a novel adverse effect of autonomy: users tend to demonstrate *initial weight inertia* - failing to learn and improve using performance feedback, driven by anchoring and ceiling effects on their initial chosen weights. Our work aligns with known behavioral biases in human-AI collaboration, such as naïve advice weighting (NAW), where people place constant weight on AI predictions because they cannot distinguish whether the impact of their private information is low or high (Balakrishnan et al. 2024). Importantly, our broader and more general setting reveals that *initial weight inertia* persists even in the absence of information asymmetry, suggesting that NAW behavior is likely to be a special case of a more fundamental behavioral pattern in human-AI decision weighting.

We also provide a potential mitigating approach and investigate its underlying mechanism to improve performance. This is particularly significant as providing full autonomy to humans can result in deviations from the optimal policy, leading to reduced productivity (Ibanez et al. 2018, Dai and Abràmoff 2023, Beer et al. 2024), revenue (Caro and de Tejada Cuenca 2023), profitability (Kesavan and Kushwaha 2020), and incurring significant costs (Kawaguchi 2021, Sun et al. 2022).

Our research contributes to the study of cognitive effort and learning in the presence of AI input. For instance, previous research has explored how AI input shapes cognitive effort (Boyaci et al. 2024), how humans' own cognitive challenges affect collaboration quality (Fügener et al. 2022), and how time pressure impacts collaboration (Snyder et al. 2024). Our research demonstrates that bounded autonomy helps users develop better strategies of integrating AI predictions by not only reducing *initial weight inertia* but also triggering greater cognitive effort in decision-making. This increased engagement is crucial, as Boyaci et al. (2024) demonstrate that cognitive effort plays a key role in accurately assessing AI quality and effective integration of human and AI inputs.

Our work also contributes to research on human-AI complementarity in decision-making (Dietvorst et al. 2018, Choudhary et al. 2025, Ibrahim et al. 2021, Balakrishnan et al. 2024, Chen et al. 2023, Snyder et al. 2024) by introducing a bounded autonomy approach that adaptively

weights AI and human predictions. This approach helps navigate the trade-off between both algorithm aversion that leads to under-reliance (Dietvorst et al. 2015) and the limitations of unbounded autonomy approaches that fail to effectively incorporate human intuition (Balakrishnan et al. 2024, Chen et al. 2023).

Our work aligns with policymakers' emphasis on human agency in AI systems, as exemplified by the EU AI Act's[2] requirements for human oversight and transparency in high-risk AI applications. The Act specifically mandates quality, transparency, and human oversight obligations for AI systems used in critical decision-making contexts. While both regulatory frameworks and industry analyses[3] emphasize preserving human autonomy in human-AI collaboration, our findings reveal a significant bias that limits optimal AI utilization. To address this issue, we propose a method to address this bias, thereby enhancing the predictive accuracy of human-AI collaborative systems.

2. Related Literature and Hypotheses Development

Given the interdisciplinary nature of AI research, our work draws from multiple streams of literature including psychology, operations management, and decision sciences. To provide a succinct review, we organize the literature into two broad categories central to our paper: (1) the role of autonomy in human-AI collaboration performance, and (2) human weighting behavior under autonomy. Together, these categories form the foundation for the development of our hypotheses.

2.1. Decision Autonomy and Human-AI Collaborative Performance

Decision autonomy — defined as the ability to independently make decisions—has emerged as a critical focus in human-AI collaboration, particularly in high-stakes contexts such as medical diagnoses and legal decisions, where accountability must remain with humans (Jussupow et al. 2021). Regulatory frameworks, such as the EU AI Act, further emphasize preserving human agency by requiring transparency and oversight in AI interactions (Green 2022).

The relationship between autonomy and decision performance is far from being straightforward. On one hand, autonomy fosters engagement and acceptance of AI systems (Dietvorst et al. 2018, Burton et al. 2020, Fink et al. 2024). On the other hand, excessive autonomy can lead to suboptimal outcomes, such as under-reliance on AI due to overconfidence in human predictions (Logg et al. 2019, Dargnies et al. 2024) or over-reliance resulting from misplaced trust in AI predictions (Sarter and Schroeder 2001, Parasuraman and Manzey 2010,

[2] https://en.wikipedia.org/wiki/Artificial_Intelligence_Act

[3] https://www.bcg.com/publications/2022/the-value-of-ai-for-individuals

Goddard et al. 2012). This duality highlights the need for structured approaches to designing autonomy that strike right balance between affording human agency while enhancing decision performance. Our study addresses this gap by examining how bounded autonomy approach — which constrains the degree of freedom in weighting AI inputs — can guide users toward more accurate integration of AI predictions.

Providing an instructional strategy under bounded autonomy that is based on past performance is likely to not only improve performance accuracy but also alleviate cognitive challenges and inattention during decision-making. These cognitive difficulties, such as evaluating AI quality, are frequently cited as obstacles in human-AI collaboration (Boyaci et al. 2024) and are particularly relevant in operations management (Bastani et al. 2021).

We hypothesize that bounding autonomy simplifies decision-making and encourages people to pay closer attention to feedback, leading to improved prediction performance. By constraining autonomy, users are encouraged to evaluate AI inputs more critically and adapt their strategies for weighting AI predictions. This approach reduces the risks of both over-reliance and under-reliance, ultimately enhancing the accuracy of human-AI collaboration. Based on this reasoning, we formally hypothesize:

HYPOTHESIS 1. *Bounded autonomy in AI prediction weighting enhances the overall human-AI collaborative prediction accuracy compared to unbounded autonomy.*

2.2. Autonomy and Initial Weight Inertia

To further contrast the effects of full autonomy and imposing bounds on it, we review relevant literature to examine how users utilize autonomy when integrating AI predictions. Our review synthesizes three research streams: (1) patterns of human-AI prediction aggregation, (2) the effects of performance feedback on learning and weighting decisions, and (3) the potential benefits of bounded autonomy. This synthesis uncovers key mechanisms that not only inform additional hypotheses but also provide theoretical support for HYPOTHESIS 1.

2.2.1. Aggregation of Human and AI Predictions Understanding how humans integrate AI predictions is fundamental to effective human-AI collaboration. Prior research has identified several integration approaches such as using ensembles of human and AI predictions in various sequences (Choudhary et al. 2025). Our research focuses on the sequential task, which is prevalent in operations management applications (Ibrahim et al. 2021) and provides humans with agency in both incorporating AI predictions and making final decisions (Logg et al. 2019, Balakrishnan et al. 2024).

Research on human-AI prediction systems follows two main paths. The first examines how AI systems can optimally incorporate human input to generate collaborative predictions (Ibrahim and

Kim 2019, Ibrahim et al. 2021, Brau et al. 2023). The second stream focuses on scenarios where humans control the final decision authority (i.e., gatekeeper) while using AI predictions as input (Önkal et al. 2009, Logg et al. 2019, Luong et al. 2020, Balakrishnan et al. 2024). Our paper aligns with the second stream where humans have the autonomy to make the final decision.

These tasks can be studied through the lens of the Judge-Advisor System (JAS) framework. In this system, humans make an original prediction, receive AI advice, and then make a final prediction incorporating this input (Bonaccio and Dalal 2006, Logg et al. 2019, Lehmann et al. 2022, Balakrishnan et al. 2024).[4] The Weight on Advice (WOA) metric quantifies how much humans rely on AI predictions by measuring where their final prediction falls between their original predictions and the AI's advice. Research shows that people often assign weight to AI advice to combine with their individual predictions (Gino and Moore 2007, Logg et al. 2019), sometimes weighing others' predictions almost equally to their own (Larrick and Soll 2006, Soll and Larrick 2009).

Balakrishnan et al. (2024) identify naïve advice weighting (NAW) as a key challenge in human-AI collaboration - where humans apply constant weights to AI predictions regardless of the impact of their own private information. This failure to appropriately adjust their weighting strategy leads to suboptimal outcomes. While prior research has documented NAW and its persistence in settings with information asymmetry between humans and AI, less is understood about the weighting pattern in a more general setting where humans and AI have access to identical information. Further, while Balakrishnan et al. (2024) propose algorithm transparency as a mitigation approach for NAW, such approaches may become insufficient in general settings where information asymmetry is not the root cause (Lehmann et al. 2022). Our research fills this critical gap by examining weight assignment behavior in this more general context and proposing an alternative mitigation approach.

2.2.2. Feedback, Learning, and Weighting Decisions In a repeated task environment, individuals received feedback on the performance of each round. Research shows that performance feedback serves as a crucial basis for evaluating one's ability to perform successfully on subsequent tasks (Bandura 1991), and this feedback transparency also shapes decision-makers' behavior through relative performance comparisons (Buell et al. 2019). Feedback has been shown to effectively improve decision-making behavior and task performance across various domains (Dahlinger et al. 2018, Gnewuch et al. 2018, Tiefenbeck et al. 2018). However, the relationship between feedback and performance improvement is contingent on context – in some cases, feedback can actually hinder performance and learning (Choudhary et al. 2021). This complexity

[4] This differs from the human-in-the-loop approach, where the human acts as a trainer for the AI. The ultimate goal of the human-in-the-loop is to automate the prediction process, with human intervention only when the AI falters.

becomes particularly salient when users must process feedback while simultaneously managing the collaborative decision-making process with AI systems.

We investigate what participants learn in our experimental setup. Participants can use feedback in two ways: (1) to make better predictions without AI input, or (2) to learn about their own and the AI's relative errors and determine better weights that improve performance through error cancellation between the two agents. This extends our understanding of the current literature.

Incorporating feedback into the subsequent decision increases cognitive load in determining appropriate weights for AI input (You et al. 2022). This cognitive burden impacts in-task learning (Sweller 1988), which also extends to effective collaboration with AI systems. Users may use feedback to assess their own accuracy, evaluate the AI's accuracy, and develop strategies for leveraging AI inputs—all of which create substantial cognitive demands.

Research has shown that when cognitive resources are heavily taxed by multiple processing requirements, users' capacity to learn and adopt optimal strategies may be impaired (Paas and Van Merriënboer 1994). Under such cognitive constraints, people typically resort to simplified heuristics to reduce cognitive demands (Tversky and Kahneman 1974). Two particularly relevant heuristics are the anchoring effect, where initial judgments disproportionately influence subsequent decisions (Epley and Gilovich 2006), and status quo bias, where people tend to maintain their existing strategies even when better alternatives exist (Samuelson and Zeckhauser 1988). Additionally, Moreover, as people become less sensitive to prediction errors over time (Dietvorst and Bharti 2020), they become increasingly reluctant to modify their decisions, even when faced with evidence of poor performance.

These simplified decision heuristics might lead to *initial weight inertia* that persists even in the face of round-by-round performance feedback, an approach that has shown limited effectiveness across various tasks (e.g., immediate feedback Choudhary et al. 2021). Therefore, we predict that in the presence of full autonomy and the lack of structured guidance, participants may show *initial weight inertia*. Formally,

Hypothesis 2. *Users exhibit initial weight inertia on AI predictions when provided with unbounded autonomy.*

Prior research has not offered concrete approaches to address such *initial weight inertia*. Therefore, we attempt to address this gap by proposing bounded autonomy with adaptive weights.

2.2.3. Bounded Autonomy and Dynamic Weighting In the absence of constraints on autonomy in a human-AI collaborative task, the *initial weight inertia* is exacerbated by the unbounded autonomy. Research suggests that individuals struggle to dynamically adjust their

weighting strategies even when presented with evidence of AI's superior predictive accuracy (Soule et al. 2023, Balakrishnan et al. 2024). Our work relates to this literature, as we incorporate feedback into our framework in the design of autonomy. We argue that restricting autonomy due to poor performance can act as a feedback signal, helping users learn—for instance, when a user overestimates their predictions. We hypothesize that by limiting permissible weighting ranges based on the predictive accuracy of AI agent and the human-AI collaboration, the bounded autonomy approach simplifies decision-making and encourages humans to move away from *initial weight inertia* – where individuals have a strong anchor based on the initial weight and fail to calibrate the weight assigned to AI predictions dynamically – thereby reducing the risk of poor accuracy outcomes. Such calibrated bounds are expected to guide users toward more effective integration of AI inputs. Thus, we propose the following hypotheses:

HYPOTHESIS 3. *Bounded autonomy in AI prediction weighting encourage users to move away from initial weight inertia, enabling them to better adapt to AI accuracy, as compared to unbounded autonomy.*[5]

3. Experiment Design

To test our hypotheses, we adopt a prediction task developed by Ibrahim et al. (2021). In this task, participants receive information about the number of procedures involved and the anesthesia complexity score for each surgery (features or x variables) and are asked to predict the duration of the surgery (outcome or y variable). In the treatment conditions, we provide participants with AI predictions on the same task and allow them to assign varying weights to these predictions under different types of autonomy when combining them with their original predictions when revising predictions. This approach enables us to test our hypothesis by comparing weights put by individuals after every round and the resulting accuracy when the autonomy is varied.

3.1. Task and Data Generation

Similar to Ibrahim et al. (2021), we simulate the AI algorithm's surgery prediction using the linear model as

$$\hat{Y}_t^{\mathrm{AI}} = 60 + 20X_t^{\mathrm{P}} + 10X_t^{\mathrm{C}} \tag{1}$$

Here, $\hat{Y}_t^{\mathrm{AI}}$ denotes the AI's duration prediction of surgery t; X_t^{P} denotes the number of procedures of surgery t, an integer-valued public factor with a uniform distribution between 1 and 10, inclusive;

[5] While we initially used the terminology of naïve weighting strategies in our pre-registration, we adopt the term initial weight inertia to distinguish our more general setting from the information asymmetry context in Balakrishnan et al. (2024). Hypotheses 1 and 3 were pre-registered. We included Hypothesis 2 post-registration to provide a more nuanced understanding of weighting behavior under unbounded autonomy.

and X_t^{C} denotes the anesthesia complexity of surgery t with a uniform distribution between -5 and 5.

The actual surgery duration (ground truth) is simulated using:

$$Y_t = 60 + 20X_t^{\mathrm{P}} + 10X_t^{\mathrm{C}} + \epsilon_t \tag{2}$$

Here, Y_t is the actual duration of surgery t; X_t^{P} and X_t^{C} denote the same as in the AI's prediction. Finally, the error term ϵ_t follows a normal distribution with mean 0 and a standard deviation of one-half the AI's prediction standard deviation of surgery t.

To understand how participants adjust their weighting strategy according to the accuracy of AI predictions, we manipulated the variation in the actual duration (ground truth) by introducing errors ranging from zero (perfect predictions) up to 100 units. Consequently, the mean absolute error (MAE) of our AI predictions was 28.5, with a standard deviation of 30.94. The median absolute deviation from the actual outcomes was 15. All participants made the same 20 decisions across conditions; however, the order of surgeries was randomized across participants to alleviate any order effect.

3.2. Conditions

Participants were randomly assigned to one of the following four conditions:

1. *No_AI Assistance (baseline).* The *No_AI* assistance condition served as our control group, where each participant i made predictions without AI support. In this condition, participants first made original predictions, denoted as $\hat{y}_{it}^{\text{original}}$. And they had an opportunity to revise these predictions without any additional information before final submission. This revision stage served as an effort control, matching the decision steps in conditions with AI assistance. The final prediction represents participants' potential revision of their original predictions, rather than a human-AI collaborative prediction, denoted as $\hat{y}_{it}^{\text{revised}}$. After submitting revised predictions, participants received performance feedback each round, allowing us to study how individuals learn from their individual performance without AI assistance. This control condition serves as a baseline for measuring the impact of AI assistance on prediction accuracy.

2. *Unbounded Autonomy in Weighting AI Predictions (UA_in_AI).* In this condition, similar to the *No_AI* assistance condition, each participant made $\hat{y}_{it}^{\text{original}}$ in the first step. They were then informed that *an AI tool (a machine learning algorithm that learns from data and makes predictions) provides predictions for this task.* Participants could freely assign any weight between 0% and 100% to the AI's predictions for each prediction task t, denoted as WOA_{it}, with the remaining percentage automatically applied to their original prediction. For instance, if a participant chose a WOA_{it} at 40%, their original prediction would receive the complementary 60% weight. Using these weights, a revised human-AI collaborative prediction was calculated as:

$$\hat{y}_{it}^{\text{revised}} = \hat{y}_{it}^{\text{original}} \times (1 - WOA_{it}) + \hat{y}_{it}^{\text{AI}} \times WOA_{it} \tag{3}$$

This condition allows us to examine how each participant naturally incorporates AI prediction for each round of prediction when given complete freedom over its utilization.

3. *Bounded Autonomy in Weighting AI Predictions (BA_in_AI).* The *BA_in_AI* condition follows a similar procedure to *UA_in_AI*, but with one key difference in the range of available weights that participants can assign to AI predictions in each round. Instead of full autonomy, participants worked within dynamic constraints when assigning weights to AI predictions. While they could still weight AI predictions up to 100%, the minimum allowable weight was adjusted adaptively based on their relative accuracy[6] - specifically, how their revised human-AI collaborative prediction compared to the AI's prediction from the previous round $t-1$. Specifically, the minimum allowable weight for each round, denoted as WOA_{it}^{min}, that participants can assign to the AI is calculated as follows:

$$WOA_{it}^{min} = \frac{|y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{revised}}|}{|y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{revised}}| + |y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{AI}}|} \times 100\% \tag{4}$$

The system dynamically calculates minimum AI weights based on relative performance in the previous round. Specifically, it compares two metrics: (1) the absolute error (AE) of the participant's revised prediction after collaboration with AI ($|y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{revised}}|$), and (2) the combined AE of both the revised prediction and the AI prediction ($|y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{revised}}| + |y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{AI}}|$). When participants demonstrated better predictive performance by achieving lower relative errors, they were rewarded with greater autonomy in subsequent rounds - WOA_{it}^{min} was lower, allowing them a wider range of permissible WOA_{it} to choose from.

4. *Bounded Autonomy in Weighting Self-Predictions (BA_in_Self).* Prior literature indicates that individuals tend to discount external advice while over-weighting their own predictions when integrating information (Yaniv and Kleinberger 2000). Therefore, we create a mirrored version of *BA_in_AI* to examine the framing effect of how participants assign weight to their original predictions versus weight to AI predictions, denoted as WOS_{it}. This condition uses the same bounded autonomy approach but with an inverse framing: instead of setting WOA_{it}^{min}, it establishes maximum weights for participants' $\hat{y}_{it}^{\text{original}}$, denoted as WOS_{it}^{max}. Participants could assign WOS_{it} ranging from 0% up to WOS_{t}^{max} when revising their original predictions. Similar to WOA_{t}^{min}, WOS_{t}^{max} is determined dynamically based on the relative accuracy of two components from the previous round $t-1$:

[6] We set the minimum weight on AI prediction due to a ceiling effect when relying on AI predictions, even with perfect AI prediction (Dietvorst and Bharti 2020).

$$WOS_{it}^{max} = (1 - \frac{|y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{revised}}|}{|y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{revised}}| + |y_{i(t-1)} - \hat{y}_{i(t-1)}^{\text{AI}}|}) \times 100\% \quad (5)$$

Thus, when participants demonstrated better predictive performance by achieving lower relative error after integrating AI predictions, they were rewarded with greater autonomy in subsequent rounds - the maximum weight on original predictions was higher, allowing them a wider range of permissible weights to choose from. Mathematically, $WOA_t^{min} = 1 - WOS_t^{max}$.

3.3. Procedure

We programmed and developed the user interface using the software in oTree (Chen et al. 2016).

The prediction task procedure was as follows:

Step 1. *Instructions and Comprehension Tests.* Participants were introduced to the prediction task and the objective of minimizing absolute prediction error. To incentivize performance, participants were rewarded based on their prediction accuracy during the forecasting task. In addition to a fixed \$4 payment for study completion, participants received a bonus calculated as:

$$Bonus_i = \$2 - \frac{\sum_{t=1}^{20} |y_{it} - \hat{y}_{it}^{\text{revised}}|}{20} \times \frac{1}{100} \quad (6)$$

They were then tested for comprehension of the concept of absolute prediction error and the bonus payment scheme tied to their prediction accuracy before proceeding to the next step, details in Figure OA.3. Note all Tables and figures in Online Appendix are prefixed with OA such as Figure OA.1 and Table OA.1.

Step 2. *Historical Surgery Review.* Participants received instructions on two variables that determine surgery duration: the type of procedure X_t^{P} and the complexity of the anesthesia X_t^{C}. They were informed that both higher procedure counts and more complex anesthesia correlate with longer surgery durations. Before proceeding, participants had to pass a three-question comprehension test to verify their understanding (refer to Figure OA.4 for the details).

Participants then examined 10 historical surgeries presented in random order, each showing the number of procedures X_t^{P}, anesthesia complexity score X_t^{C}, and actual duration Y_t. For each surgery, participants had to enter a predicted duration before the actual value was revealed. They were informed that these training predictions would not affect their payment. After reviewing all 10 cases, participants rated the task's difficulty.

Step 3. *Actual Prediction Task.* Participants were tasked with predicting the duration of 20 new surgeries that were not previously encountered during the training phase. Before beginning these predictions, participants received non-incentivized interface instructions specific to their assigned

experimental condition. After familiarization with the interface, participants completed 20 official prediction rounds that determined their bonus payment.

The prediction procedure consisted of two stages. In the first stage, all participants generated an original prediction without any assistance $\hat{y}_{it}^{\text{original}}$, using the same interface as in the training phase (refer to Figure OA.7 for the details). In the second stage, participants could revise their original predictions $\hat{y}it^{\text{revised}}$, with or without AI assistance, based on their experimental condition. The revision stage interface varied across four conditions. In the no AI-assisted baseline condition, participants were simply given the opportunity to revise their original prediction $\hat{y}it^{\text{original}}$ to a revised prediction $\hat{y}_{it}^{\text{revised}}$, which would determine their bonus. In the AI-assisted conditions, participants assigned weights to either AI predictions (WOA_{it}) or original predictions (WOS_{it}) for each round. (refer to Figures OA.8 through OA.17 for the details).

Surgery predictions were presented in random order for each participant, with an attention check administered at the midpoint of these prediction rounds.

Performance Feedback After submitting revised predictions, participants received feedback each round showing: the actual surgery duration y_{it}, their revised final prediction (with or without AI assistance) $\hat{y}_{it}^{\text{revised}}$, and the absolute error between these values $|y_{it} - \hat{y}_{it}^{\text{revised}}|$.

Step 4. *Survey.* Subsequently, all participants completed survey questions to assess the perceived difficulty of the prediction task and their confidence in their predictions. Participants assigned to conditions involving AI assistance also answered additional survey questions. These questions measured trust in the AI, perceptions of the AI's accuracy, beliefs about mistakes made by themselves compared to the AI algorithm, perceived control over the task, and prior experience with machine learning and generative AI models such as ChatGPT or other large language models (LLMs).

Step 5. *Payment.* Participants were provided feedback on their prediction accuracy, measured as the average absolute error across the 20 prediction rounds, as well as their bonus payment amount based on that prediction accuracy. They were also provided with a link to return to Prolific for completion.

We did not collect any personally identifiable information to protect the privacy of the participants. Participants were debriefed upon study completion.

The detailed step-by-step screenshots of the experimental procedure are provided in Online Appendix 1.

3.4. Pre-registration and Data Availability

We pre-registered the experimental details on `https://aspredicted.org/TMG_XJM`. Our pre-registration specified the target sample size, exclusion criteria, and planned analyses. We

pre-registered to exclude participants who (1) predicted the same values for 90% or higher of the predictions or (2) who did not provide answers to all the questions. The key dependent variable was prediction error, which we calculated as the absolute difference between participants' revised predictions and true values (ground truths). Detailed experiment materials, data, and data analysis were made available through an anonymized link at Open Science Framework.[7]

4. Data and Key Variables

4.1. Participant Recruitment

A total of 364 participants recruited through Prolific completed the full study procedures. However, 4 participants were excluded from the final sample as they failed to provide the correct Prolific completion code,[8] This left us with 360 participants (176 males, 180 females, and 4 participants with undisclosed gender) for our analysis meeting our preregistered sample size.[9] After randomly assignment across four conditions, we had 91 participants in the *No_AI* condition, 89 in the condition of *UA_in_AI*, 89 in the condition of *BA_in_AI*, and 91 in the condition of *BA_in_Self*. On average, participants completed the study in 22.08 (SD = 12.25) minutes and received an additional bonus payment of $1.60 (SD = 0.15) based on the bonus scheme detailed in Equation 6. The demographic characteristics of participants are summarized in Table OA.1.

4.2. Main Variables and Summary Statistics

To understand how participants integrate AI predictions with their original predictions, we define the key performance measures on the data obtained for each of the 20 rounds of the actual prediction task in Table 1. The first three variables are measurement of prediction performance whereas the last two variables measure weight on AI advice.

Our sample consists of 1,820 observations for the baseline group (*No_AI*) and 5,380 observations for the AI-assisted groups (comprising *UA_in_AI*, *BA_in_AI*, and *BA_in_Self*). While the average *AE_original* is similar between the baseline (57.68) and AI-assisted groups (56.70), there is a notable difference in their maximum values. The AI-assisted groups show a substantially lower maximum *AE_original* of 471 compared to 1,559 in the baseline group, suggesting that AI assistance may help mitigate extreme prediction errors.

[7] https://osf.io/ezf67/?view˙only=f2294197eccb4651a9efd76cf52cfbca

[8] Of the excluded participants, one provided an incorrect completion code, one with an unknown completion code, and two experienced technical timeouts on Prolific. To maintain fairness towards participants recruited through Prolific, these individuals were compensated fully for their time spent, including bonus payment.

[9] The sample size of 360 was preregistered based on a medium effect size (d = 0.4), alpha = 0.05, and power = 0.8, requiring 312 participants. Extra participants were added for potential exclusions.

Table 1: Description of the Main Variables

Variable	Description
AE_original	Absolute error of participants' original prediction in each round
AE_revised	Absolute error of participants' revised final prediction in each round (incorporating AI advice for the AI-assisted groups)
AE_AI	Absolute error of AI's prediction in each round.
WOA	Participants' assigned weight to integrate AI predictions with their own (%).
WOA_initial	Participants' assigned weight to integrate AI predictions with their own during the first round (%).

The analysis of *AE_revised* reveals more pronounced differences between groups. The AI-assisted groups show a lower average *AE_revised* of 34.53 compared to 57.30 in the baseline group, with their maximum error (292) also substantially lower than the baseline group's maximum of 1,302. For the AI-assisted groups, the average *WOA* is 66.29, ranging from 0 to 100, with *WOA_initial* averaging 56.47. The *AE_AI* averages 28.50, with a maximum of 100, suggesting that AI predictions are generally more accurate than the original predictions.

Table 2: Summary Statistics of Main Variables

	Group	Min	Mean (SD)	Max	N
AE_original	*No_AI*	0	57.68 (61.26)	1559	1820
	AI-assisted	0	56.70 (49.68)	471	5380
AE_revised	*No_AI*	0	57.30 (57.89)	1302	1820
	AI-assisted	0	34.53 (35.43)	292	5380
WOA	AI-assisted	0	66.29 (25.22)	100	5380
WOA_initial	AI-assisted	0	56.47 (28.41)	100	89
AE_AI	AI-assisted	0	28.50 (30.94)	100	5380

Note that all participants were provided with identical information during the practice rounds[10]. We do not find any significant differences in practice performance among the four experimental conditions (refer to Table OA.2 for details), confirming the similar baseline across groups due to randomization. In the following section, we examine these patterns in greater detail by analyzing how different groups leverage AI assistance and how it impacts their performance.

[10] The practice rounds were not incentivized; participants could enter any number they wished, as these rounds served solely for familiarization with the task.

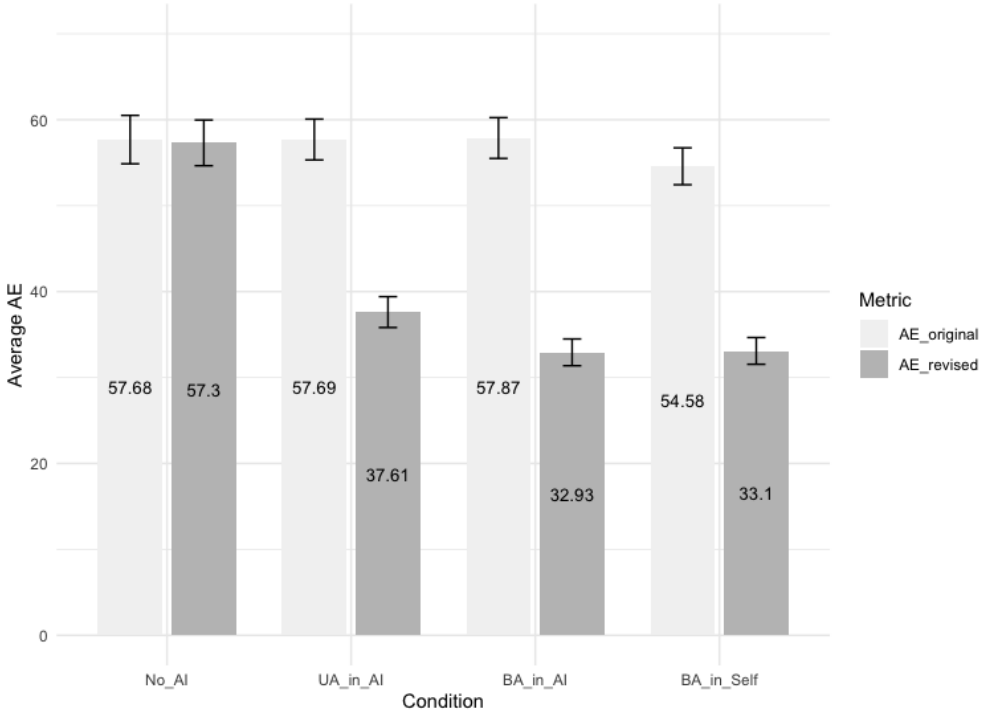
5. Results

In this section, we present our empirical findings by examining three key aspects: prediction accuracy, *initial weight inertia*, and the impact of bounded autonomy on weighting behaviors. Our analysis proceeds as follows: first, we examine model-free descriptive evidence of prediction accuracy followed by a model estimation to test whether bounded autonomy improves overall performance. Second, we investigate *initial weight inertia* under bounded autonomy by a) examining the anchoring and ceiling effects in *WOA_initial*, b) assessing the responsiveness of *WOA* to performance feedback, and c) evaluating whether this *initial weight inertia* leads to deviations from optimal weights. Third, we evaluate whether bounded autonomy mitigates *initial weight inertia*. Fourth, we compare weighting behaviors under different framing conditions, to provide support for our proposed mechanism. Finally, we synthesize these findings to offer practical insights for managers.

5.1. Impact of Bounded Autonomy on *AE* and *WOA*

Figure 1 presents the model-free evidence of participants' prediction performance, comparing both *AE_original* and *AE_revised* across all four conditions. The results indicate that participants' baseline prediction accuracy remains consistent, irrespective of whether they receive AI assistance (comparing *No_AI* and *UA_in_AI* conditions) or the level of autonomy granted (comparing *UA_in_AI* and *BA_in_AI/Self*). Notably, the *AE_original* in each condition is, on average, less accurate than the AI's standalone predictions, underscoring the value of incorporating AI predictions to enhance the accuracy of revised collaborative predictions.

Figure 1: *AE_orginal* vs. *AE_revised* by Condition



Note: Error bars indicate 95% confidence intervals.

Comparing *AE_original* across groups using a pairwise *t-test*, we find no statistically significant differences between them. This indicates that participants started with a similar level of prediction accuracy. While the *No_AI* group shows no significant changes between *AE_original* and *AE_revised*, the AI-assisted groups demonstrate significantly lower *AE_original* compared to their *AE_revised* (all *t-tests*, $p < 0.01$). To estimate the impact of BA on prediction error, we estimate the following model using OLS:

$$AE_{it} = \boldsymbol{\beta}\, condition_i + \varepsilon_{it} \tag{7}$$

where i represents an individual participant, t denotes the round, and ϵ is the error term. $condition_i$ denotes which condition a participant has been assigned to. The dependent variable AE represents either *AE_original* or *AE_revised*. We estimate the β coefficients to identify the impact of BA on AE. We cluster the standard errors at the individual level to account for within-participant correlation. We report the results of our estimation in Table 3. In column 1, using the *No_AI* condition as the baseline, we find no significant differences between conditions in participants' original predictions. All coefficients in column 1 are statistically insignificant, indicating that participants' original prediction accuracy is similar across conditions regardless of

the AI assistance they later receive. The estimate *Constant* (57.68) represents the average *AE_original* for the *No_AI* group.

Column 2 shows that participants in the *UA_in_AI* condition significantly reduce their *AE_revised* by 34.36% after receiving AI assistance (-19.69, $p < 0.01$). Similarly significant reductions are observed for *BA_in_AI* (-24.38, $p < 0.01$) and *BA_in_Self* (-24.21, $p < 0.01$), representing improvements of 42.54% and 42.25% respectively. These results demonstrate that AI assistance substantially improves prediction accuracy compared to the baseline condition.

Table 3: Impact of BA on *AE_orginal* and *AE_revised*

	AE_original	*AE_revised*
	(1)	(2)
UA_in_AI	0.01	-19.69***
	(2.76)	(2.32)
BA_in_AI	0.18	-24.38***
	(2.87)	(2.01)
BA_in_Self	-3.10	-24.21***
	(2.54)	(2.03)
Constant	57.68***	57.30***
	(2.02)	(1.98)
N	7200	7200
Adjusted R^2	0.00	0.05

*Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors are in parentheses and clustered at an individual level.*

We next study the impact of bounded autonomy on *AE* compared to *UA_in_AI* condition. Comparing the coefficients of *AE_revised* for *BA_in_AI* (-24.38, $p < 0.01$) and *UA_in_AI* (-19.69, $p < 0.01$) reveals they are statistically different ($p < 0.01$)[11]. The *BA_in_AI* condition achieves an average *AE_revised* of 32.92 compared to 37.61 in the *UA_in_AI*, representing a 12.47% reduction in *AE*. The coefficients of *BA_in_AI* (-24.38, $p < 0.01$) and *BA_in_Self* (-24.21, $p < 0.01$) are not statistically different. This indicates both implementations of the bounded autonomy approach improve human-AI collaborative prediction performance. These results support our HYPOTHESIS 1, which predicted that bounded autonomy enhances the overall accuracy of participants' predictions compared to unbounded autonomy.

[11] We conduct pairwise comparisons of treatment effects using linear hypothesis tests.

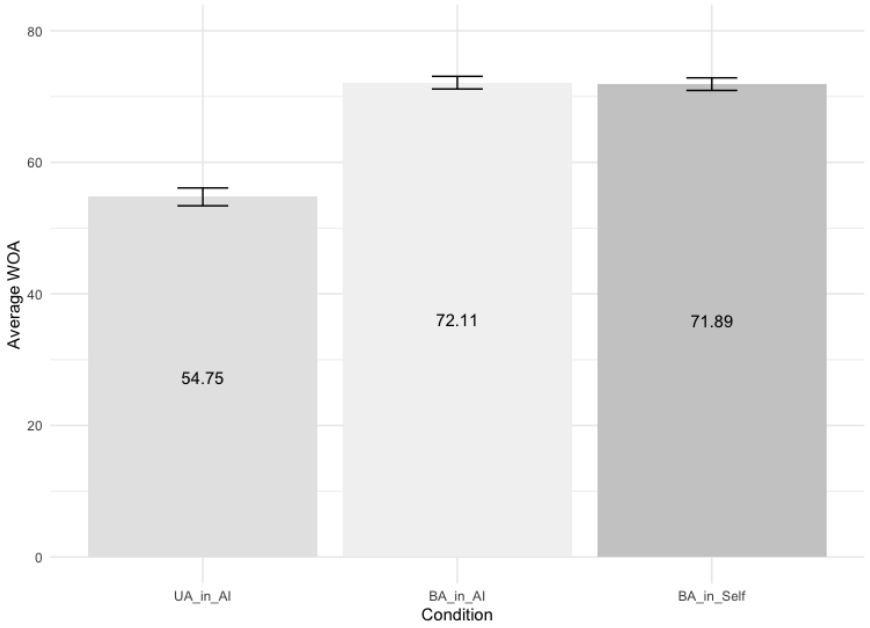
5.1.1. Does Range Reduction Mechanically Drive the Results? One alternative explanation is that the performance improvements stem from mechanical constraints on weight ranges rather than participants' improved use of AI input. We calculate post-hoc optimal weights per round based on participants' original predictions to determine both minimum and maximum AE bounds for AI-assisted conditions (refer to Online Appendix 2 for detailed calculations of optimal weights). The *maximum AE* in *UA_in_AI* occurs at either 0% or 100% AI reliance, while in *BA_in_AI* it occurs at either minimum weight or 100% AI reliance. The average [*minimum AE*, *maximum AE*] ranges are [19.78, 62.36] for *UA_in_AI*, [23.04, 43.37] for *BA_in_AI*, and [23.00, 41.93] for *BA_in_Self* (refer to Table OA.3 for the details). These ranges indicates that while BA conditions reduce extreme errors by preventing extreme weights, they also limit the potential for achieving optimal accuracy that is attainable under UA. Thus, merely bounding the weights does not mechanically lead to improvement in accuracy, supporting the conclusion that participants make better decisions when subjected to bounded autonomy.

We also conduct a counterfactual analysis by calculating the optimal weights for participants under bounded autonomy if their weights were not bounded. This analysis reveals that in 26.46% of *BA_in_AI* cases and 27.53% of *BA_in_Self* cases, the mathematically optimal weights fall below the minimum weight constrained by the BA conditions. This demonstrates that the minimum weight requirement in BA conditions can actively prevent participants from achieving better accuracy that would be possible under the *UA_in_AI* condition, unless they carefully consider their weighting decisions. Thus, merely bounding the weights does not always lead to improvement in accuracy.

5.1.2. How Do Participants Assign *WOA* Across Conditions? Having established that bounded autonomy systematically improves performance, we examine how participants assign *WOA*, as shown in Figure 2. Participants assigned in the *UA_in_AI* condition assign an average 54.75% (SD = 29.20) *WOA*, which is significantly different from 50% ($p < 0.01$), albeit close to the midpoint between the original and AI predictions.

In the *BA_in_AI* condition, the average *WOA* is 72.11% (SD = 20.70), a 31.70% increase compared to *UA_in_AI* ($t = 7.26$, $p < 0.01$). A similar pattern emerges in *BA_in_Self*, where participants assign an average *WOA* of 71.89% (SD = 20.81), representing a 31.31% increase compared to *UA_in_AI* ($t = 7.44$, $p < 0.01$). No significant difference is observed between *BA_in_AI* and *BA_in_Self*. The higher weights and lower standard deviations under bounded autonomy conditions indicate that this condition leads to both greater and more consistent reliance on AI predictions. In the following sections, we compare *WOA* across *UA_in_AI*, *BA_in_AI* and *BA_in_Self* in details. We begin by investigating *WOA* under *UA_in_AI* in details.

Figure 2: *WOA* by Condition



Note: Error bars indicate 95% confidence intervals.

5.2. Understanding Initial Weight Inertia Under *UA_in_AI*

To understand *initial weight inertia* under *UA_in_AI*, we examine how *WOA_initial* predicts subsequent weight adjustments and assess the responsiveness of *WOA* to performance feedback.

To determine whether *WOA_initial* serves as a reference point that most participants anchor to, we conduct a fixed-effects linear regression to assess how *WOA_initial* associates with subsequent weighting decisions. As shown in Table 4, the coefficient of *WOA_initial* in column 1 is positive and significant albeit with a small coefficient (0.40, $p < 0.01$). This suggests that under *UA_in_AI*, *WOA_initial* acts as an anchor in subsequent weight adjustments.

An alternative explanation is that participants learn about weights over rounds and adjust their weights compared to the previous round. If this holds, controlling for the lagged *WOA* should make the coefficient of *WOA_initial* insignificant. We find no such evidence in column 2, as *WOA_initial* remains a positive and significant predictor (0.20, $p < 0.01$), and the same applies to *lag_WOA* (0.48, $p < 0.01$). These findings demonstrate two key patterns of anchoring: first, participants' initial weighting decisions create lasting anchors that are associated with their future weighting decisions. Second, the weight assigned in each round is significantly associated with the next round's decisions (higher *WOA* assignments tend to persist in subsequent rounds). This dual evidence of both initial anchoring and round-to-round persistence indicates that participants exhibit strong inertia in their weighting strategies.

Table 4: Anchoring Effects of *WOA_initial* Across Rounds Under *UA_in_AI*

	WOA *(excluding WOA_initial)*	
	(1)	(2)
WOA_initial	0.40***	0.20***
	(0.08)	(0.05)
lag_WOA		0.48***
		(0.05)
Constant	30.65***	16.71***
	(5.00)	(3.23)
N	1691	1691
Adjusted R^2	0.15	0.34

*Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors are in parentheses and clustered at an individual level.*

Next, we plot the round-by-round average *WOA*. Figure 3 presents the average *WOA* for each round under the *UA_in_AI* condition. Participants initially assign an average weight of 59.65% (SD = 28.51) to AI predictions. Notably, despite having unbounded autonomy to adjust weights between 0% and 100% in every round, participants' reliance on AI predictions does not exceed their *WOA_initial* in subsequent rounds, suggesting a ceiling effect on AI weights. Analysis of participants' maximum *WOA* across the 20 prediction rounds under *UA_in_AI* reveals that for ˜18% of participants, their maximum *WOA* across all 20 rounds remains at *WOA_initial*. As shown in Figure 4, this percentage is significantly higher than both *BA_in_AI* and *BA_in_Self*. Thus, *UA_in_AI* exhibits a significant ceiling effect which diminishes once the autonomy is bounded.

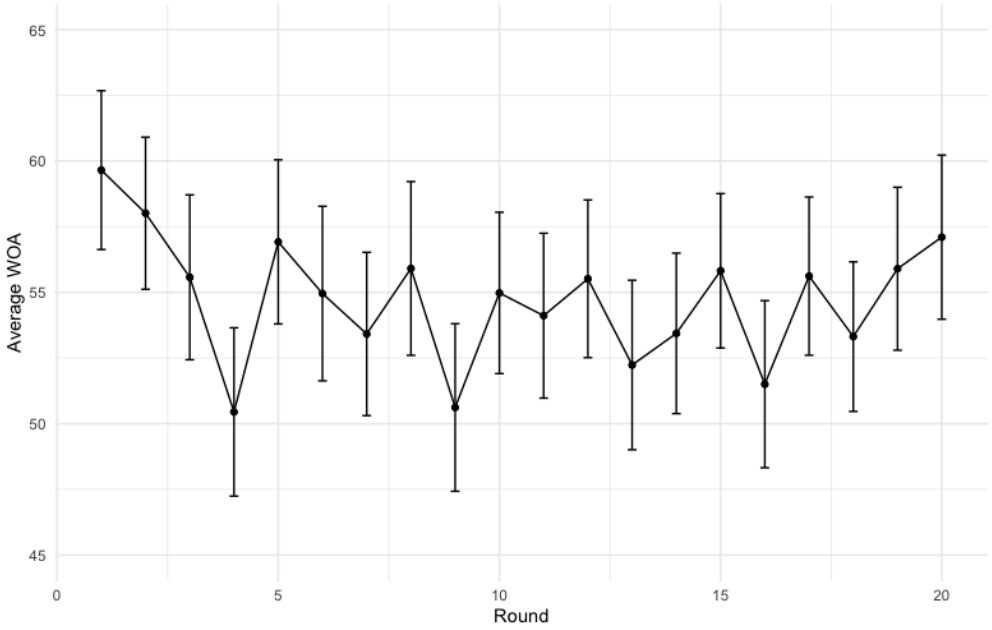
5.2.1. Do Participants Attend to Performance Feedback Under *UA_in_AI*? To further examine the *initial weight inertia*, we analyze whether participants adjust their *WOA* in response to performance feedback. Lack of adjustment would support the *initial weight inertia* effect. Using a fixed-effect linear regression model, we analyze how participants adjust their AI weighting based on performance feedback from the previous round using the following model.[12]

$$WOA_{it} = \beta\, AE\_revised_{i(t-1)} + fe_i + \varepsilon_{it} \tag{8}$$

In equation 8, the dependent variable WOA_t represents the weight assigned to AI prediction for each round. The independent variable $AE\_revised_{t-1}$, represents the lagged AE of revised predictions after collaboration with AI. The model includes participant fixed effects (fe_i) to control for time-invariant individual characteristics. Table 5 summarizes the results.

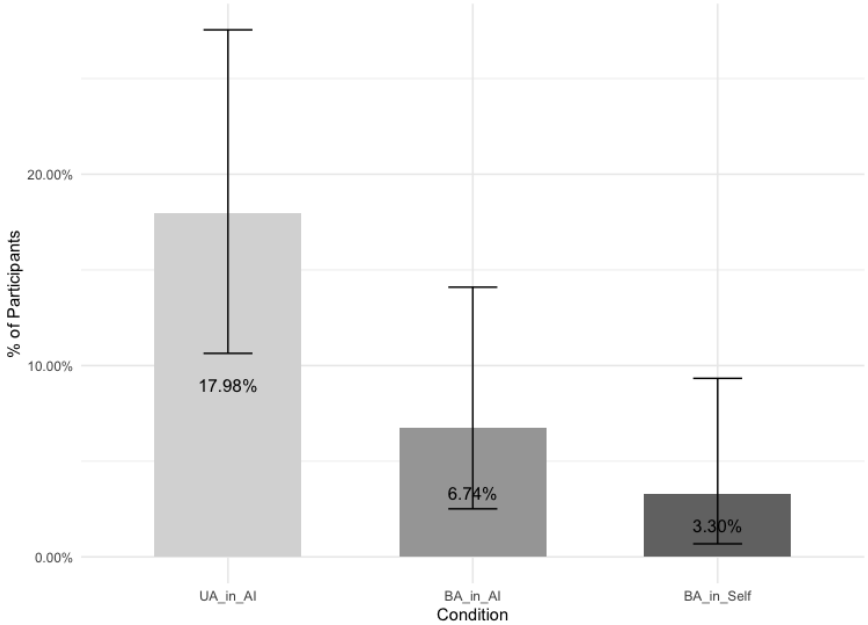
[12] Following the recency effect (Greene 1986), we focus on the immediately preceding round's feedback, as participants are more likely to recall and respond to their most recent performance experience.

Figure 3: Average *WOA* Across Rounds Under *UA_in_AI*



Note: Error bars indicate 95% confidence intervals.

Figure 4: Proportion of Participants Who Placed Maximum *WOA* in the First Round



Note: Error bars indicate 95% confidence intervals.

We find that coefficients of lagged *AE_revised* are not significant, even after controlling for AE_AI_{t-1} (lagged AE of AI predictions). These results indicate that participants do not

Table 5: *WOA* Adjustment Based on Performance Feedback Under *UA_in_AI*

	WOA	
	(1)	(2)
lag_AE_revised	−0.02	-0.01
	(0.02)	0.03
lag_AE_AI		−0.02
		(0.03)
Fixed-effects Individual ID	Yes	Yes
N	1691	1691
Adjusted R^2	0.47	0.47

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Standard errors are in parentheses and clustered at an individual level, estimated with individual participant fixed effects.

appropriately adjust their *WOA* based on performance feedback. Rational behavior predicts that participants, as sophisticated decision-makers, would modify their *WOA* each round according to comparative performance. Specifically, when lagged *AE_revised* increases (suggesting that revised predictions after collaboration with AI performed poorly), participants should logically increase their *WOA* based on the assumption that their *AE_original* is higher than *AE_AI*. However, we find no significant relationship between lagged *AE_revised* either with or without controlling for lagged *AE_AI*, and *WOA*. These results suggest that participants maintain relatively constant weighting strategies in the presence of feedback rather than optimally updating them based on relative prediction accuracy, supporting the *initial weight inertia*.

The unresponsiveness to performance feedback, combined with the persistence of anchoring and ceiling effects of initial weights on subsequent weighting decisions, demonstrates the *initial weight inertia* effect, supporting HYPOTHESIS 2.

To understand whether *initial weight inertia* leads to higher prediction errors in unbounded autonomy (shown in Figure 1), we compare *WOA* with post-hoc calculated optimal weights for each round by examining their absolute deviations. As shown in Table OA.5, participants in the *UA_in_AI* condition deviate by 40.73 points (SD = 29.16) from the average optimal weight of 72.07% (refer to Table OA.4 for details), indicating significantly suboptimal weighting behavior under unbounded autonomy.

Our analysis of weighting behavior under unbounded autonomy reveals three key patterns: first, the *initial weight inertia*, manifested as strong anchoring effects of initial weights on subsequent

weighting decisions, and the ceiling effect of placing the maximum weight on AI predictions in the first round. Second, participants dose not base their weight adjustments on the performance feedback. Third, participants show significant deviation from optimal weighting by consistently underweighting AI predictions relative to their superior accuracy. These findings suggest that simply giving users complete freedom in AI weighting decisions may not lead to optimal human-AI collaboration. Building on these insights, we next demonstrate how bounded autonomy can effectively address these behavioral limitations and guide users toward more appropriate weighting strategies.

These findings extend prior research by Balakrishnan et al. (2024), which showed that transparency can improve performance when users have private information unavailable to AI. In a broader setting, our results demonstrate that transparency alone is insufficient, as even with full transparency, people may not assign optimal weights to AI predictions. To address this limitation, we propose *BA_in_AI* as a mitigation approach, which we investigate in detail in the next section, and have demonstrated to improve prediction accuracy.

5.3. *BA_in_AI*: Addressing Initial Weight Inertia

Although we have established that bounded autonomy improves overall performance, we now identify its underlying mechanisms and provide empirical support. We show that this improvement stems from reduced *initial weight inertia*, specifically through lower anchoring and ceiling effects, and is evidenced by increased responsiveness to performance feedback rather than merely mechanical constraints from the bounded range. To investigate these mechanisms, we analyze the *BA_in_AI* condition, as both BA conditions yield similar improvements in accuracy. Later, we analyze how these mechanisms are affected by changing the framing in *BA_in_self* condition.

5.3.1. Underlying Mechanism Similar to our analysis of *UA_in_AI*, we first test the anchoring effect of *WOA_initial* on subsequent weighting decisions under *BA_in_AI*, presented in Table 6. The coefficient of *WOA_initial* remains positive and significant (0.15, $p < 0.01$) under *BA_in_AI* but is smaller compared to *UA_in_AI* (0.40, $p < 0.01$) in columns 1 and 3. Additionally, when controlling for the lagged *WOA* in columns 2 and 4, the coefficient of *WOA_initial* significantly decreases compared to its value under *UA_in_AI* (0.14, $p < 0.01$ vs. 0.20, $p < 0.01$). However, the most prominent reduction is observed in the coefficient of *lag_WOA* in columns 2 and 4, which decreases substantially in *BA_in_AI* (0.09, $p < 0.01$) compared to the value in *UA_in_AI* (0.48, $p < 0.01$). These results indicate that although both *WOA* initial and sequential anchoring effects continue to predict subsequent weighting decisions, these effects are significantly attenuated under *BA_in_AI*. This suggests that *BA_in_AI* effectively reduces

participants' tendency to anchor their subsequent weighting decisions on their initial weights assigned to AI predictions.

Table 6: Anchoring Effects of *WOA_initial* Across Rounds (*UA_in_AI* vs *BA_in_AI*)

	WOA *(excluding WOA_initial)*			
	UA_in_AI		*BA_in_AI*	
	(1)	(2)	(3)	(4)
WOA_initial	0.40^{***}	0.20^{***}	0.15^{***}	0.14^{***}
	(0.08)	(0.05)	(0.04)	(0.04)
lag_WOA		0.48^{***}		0.09^{**}
		(0.05)		(0.04)
Constant	30.65^{***}	16.71^{***}	64.28^{***}	58.90^{***}
	(5.00)	(3.23)	(2.44)	(2.76)
N	1691	1691	1691	1691
Adjusted R^2	0.15	0.34	0.04	0.05

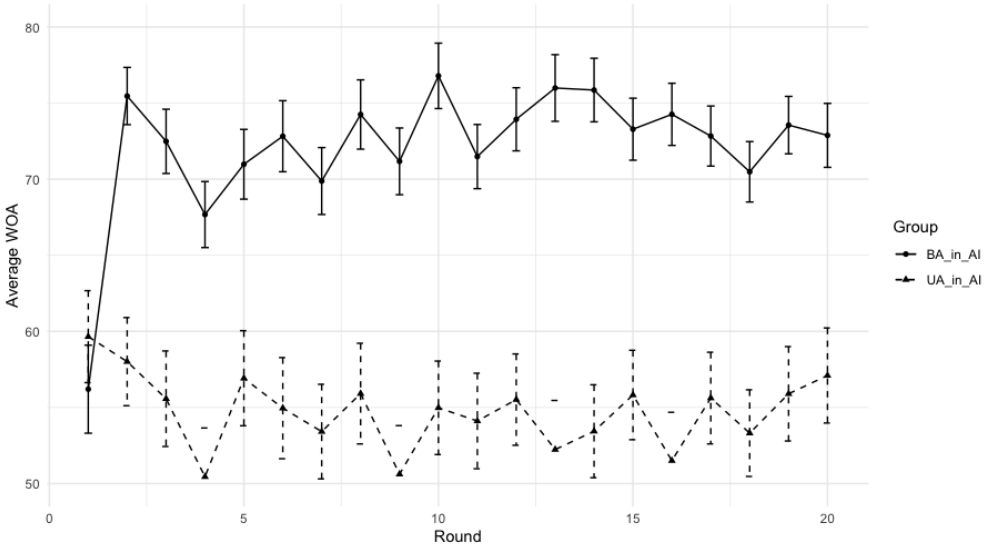
Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Standard errors are in parentheses and clustered at an individual level.

Next, we plot the round-by-round average *WOA* under *BA_in_AI* and compare it to the pattern of weighting behavior under *UA_in_AI*. In Figure 5, we observe that in the first round, average *WOA_initial* is 56.20% (SD = 29.37), not significantly different from the *WOA_initial* under *UA_in_AI*, as expected. However, the *BA_in_AI* diminishes the ceiling effect previously observed in the *UA_in_AI*. This is evident from Figure 4, as the proportion of participants who place their maximum *WOA* in the first round drops significantly from ˜18% under *UA_in_AI* to ˜7% under *BA_in_AI* (t = -2.30, $p < 0.05$).

To provide further evidence of the reduced *initial weight inertia*, we examine how participants respond to performance feedback in making *WOA* adjustments by estimating Equation 8 and reporting the results in Table 7.

The lagged *AE_revised* alone shows a negative and significant effect (-0.03, $p < 0.05$) in column 1. However, when controlling for lagged *AE_AI*, its coefficient becomes positive and significant (0.48, $p < 0.01$) in column 2. This indicates that, after controlling for AI prediction accuracy in the previous round, participants assign higher weights to AI predictions when their revised predictions are less accurate in the previous round. Lagged *AE_AI* demonstrates a significant negative effect with a larger magnitude than lagged *AE_revised* (-0.61, $p < 0.01$ vs. 0.48, $p < 0.01$), aligning with

Figure 5: Average *WOA* Across Rounds (*UA_in_AI* vs. *BA_in_AI*)



Note: Error bars indicate 95% confidence intervals.

Table 7: How Participants Adjust *WOA* Based on Performance Feedback Under *BA_in_AI*

	WOA	
	BA_in_AI	
	(1)	(2)
lag_AE_revised	-0.03^{**}	0.48^{***}
	(0.01)	(0.05)
lag_AE_AI		-0.61^{***}
		(0.05)
Fixed-effects Individual ID	Yes	Yes
N	1691	1691
Adjusted R^2	0.18	0.39

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Standard errors are in parentheses and clustered at an individual level, estimated with individual participant fixed effects.

algorithm aversion literature which shows that people lose confidence in AI more rapidly than in humans when observing AI errors (Dietvorst et al. 2015).

In summary, our results show that bounded autonomy shifts participants away from *initial weight inertia* by reducing both the anchoring and ceiling effects of *WOA_initial* on subsequent weighting decisions and increasing responsiveness to performance feedback, supporting HYPOTHESIS 3.

5.3.2. Does *BA_in_AI* Lead to Better Weighting Decisions? To answer this question, we analyze the weight deviation from the optimal weights, as previously discussed under *UA_in_AI*. Referring to Table OA.5, the results reveal that the *BA_in_AI* condition results in significantly lower deviations from optimal weights, averaging 18.05 points (SD = 20.78) compared to 40.74 points (SD = 29.16) in the *UA_in_AI* condition ($t = -13.15$, $p < 0.01$), representing a substantial reduction in suboptimal weighting behavior in *BA_in_AI* compared to *UA_in_AI*.

Our analysis of the *BA_in_AI* condition reveals three key patterns that contrast with unbounded autonomy: First, participants show reduced *initial weight inertia*, with less anchoring on their initial weights and fewer ceiling effects in early rounds. Second, participants demonstrate dynamic weighting strategies across rounds, actively adjusting their weights in response to performance feedback. Third, participants achieve more optimal weighting by assigning weights that better align with AI's demonstrated accuracy. These findings suggest that constraining users' decision space through bounded autonomy can effectively guide human-AI collaboration. Building on these insights, we next investigate whether these improved weighting behaviors persist when the decision framing changes from weighting AI predictions to weighting one's original predictions, as different psychological processes may impact decision-making despite the mathematical equivalence of these framings(Yaniv and Kleinberger 2000). This is important to understand, as policymakers can employ bounded autonomy in different ways.

5.4. Framing Effects: Weighting on AI versus Original Predictions

5.4.1. Similarities and Differences Between *BA_in_AI* and *BA_in_Self* Our analysis reveals *AE_revised* in both *BA_in_AI* and *BA_in_Self* is remarkably consistent in their performance, suggesting the equal effectiveness of these approaches in improving prediction task performance.

Despite these similarities, a deeper analysis of participants' *initial weight inertia* is needed through two key aspects: the anchoring effect of *WOA_initial* on subsequent weighting strategy, and the ceiling effect of *WOA_initial*, along with responsiveness to performance feedback.

We first employ the same estimation approach to examine the anchoring effects under *BA_in_Self* compared to *BA_in_AI* conditions (Table 8). We provide columns 1 and 2 as a reference presented earlier. The coefficients of *WOA_initial* and lagged *WOA* are both insignificant at the $p < 0.05$ level under *BA_in_Self* in columns 3 and 4. Therefore, although *BA_in_AI* and *BA_in_Self* both result in similar accuracy gains, the anchoring effect is reduced in *BA_in_AI* but is almost eliminated in *BA_in_Self*.

Similar to *BA_in_AI*, *BA_in_Self* results in a decrease of ceiling effect (refer to Figure 4 for details).

Table 8: Anchoring Effects of *WOA_initial* Across Rounds (*BA_in_AI* vs. *BA_in_Self*)

	WOA (excluding WOA_initial)			
	BA_in_AI		*BA_in_Self*	
	(1)	(2)	(3)	(4)
WOA_initial	0.15***	0.14***	0.06*	0.05
	(0.04)	(0.04)	(0.03)	(0.03)
lag_WOA		0.09**		0.06*
		(0.04)		(0.03)
N	1691	1691	1729	1729
Adjusted R^2	0.04	0.05	0.01	0.01

*Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors are in parentheses and clustered at an individual level.*

We then examine how participants respond to performance feedback under *BA_in_Self* by applying the same analysis used for *BA_in_AI*. Referring to Table OA.6, the results show that participants under *BA_in_Self* exhibit similar adjustment patterns to those in *BA_in_AI*.

However, the findings present an intriguing puzzle: while *BA_in_Self* eliminates the anchoring effect of the initial weights and maintains responsiveness to feedback similar to *BA_in_AI*, it does not yield the expected further improvements in prediction accuracy relative to *BA_in_AI*. To understand this unexpected result, we examine the underlying reasons by analyzing participants' learning patterns and cognitive effort in the following section. We incorporate both the *No_AI* and *UA_in_AI* conditions into this examination to provide a comprehensive understanding of learning dynamics and cognitive effort across all conditions.

6. Learning and Cognitive Processing in Human-AI Collaboration

Drawing from the literature on learning and cognitive effort discussed in our review, we examine how participants learned from performance feedback over time and engaged cognitively when integrating AI predictions into their decision-making process.

6.1. Impact of AI on Learning Patterns

We first investigate how AI assistance affects participants' learning across prediction rounds in Table 9. The coefficient of *AE_original* is negative and significant only under *BA_in_AI* (-0.31, $p < 0.05$) in column 3. Similarly, the coefficient of *AE_revised* is negative and significant only under *BA_in_AI* (-0.52, $p < 0.05$) in column 5. These results indicate only participants in the *BA_in_AI* condition significantly improved their original predictions over rounds. However, significant improvement in revised predictions across rounds is observed only in the *No_AI* condition, and is non-significant in all AI-assisted conditions.

Table 9: Learning in Original and Collaborative Prediction Accuracy Across Rounds

	AE_original				*AE_revised*			
	No_AI	*UA_in_AI*	*BA_in_AI*	*BA_in_Self*	*No_AI*	*UA_in_AI*	*BA_in_AI*	*BA_in_Self*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Round	-0.31^{*}	-0.09	-0.31^{**}	0.03	-0.52^{**}	-0.20	-0.28^{*}	-0.14
	(0.22)	(0.22)	(0.23)	(0.21)	(0.20)	(0.15)	(0.23)	(0.21)
Fixed Effect Individual ID	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,820	1,780	1,780	1,820	1,820	1,780	1,780	1,820
Adjusted R^2	0.05	0.07	0.10	0.05	0.06	0.04	-0.04	-0.04

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Standard errors are in parentheses and clustered at an individual level, estimated with individual participant fixed effects.

The pattern reveals an intriguing paradox in the *BA_in_AI* condition: while participants improved their original prediction accuracy over rounds, this learning did not translate into more accurate revised predictions when collaborating with AI. This suggests that even as participants became more accurate at making original predictions, they may not have effectively integrated AI's input into their improved original predictions. Conversely, in the *No_AI* condition, where participants relied solely on their own predictions, they demonstrated significant improvement in their revised predictions over time.

This pattern highlights an important trade-off: while AI assistance may enhance immediate prediction accuracy, it appears to impair participants' ability to learn and improve their collaborative predictions over time. However, this finding does not explain why the two framings of bounded autonomy (*BA_in_AI* and *BA_in_Self)* yield similar prediction accuracy despite the elimination of the anchoring effect on initial weights under *BA_in_Self*. Therefore, we next examine participants' cognitive effort during the prediction process to identify potential reasons.

6.2. Comparing Cognitive Effort Across Conditions

We examine cognitive processing by measuring processing time (PT) in seconds across three task phases: original prediction, weight assignment for AI prediction original predictions, and feedback review. We specifically focus on comparing the PT during weight assignment whether applied to AI predictions (*BA_in_AI*, *UA_in_AI*) or participants' original predictions (*BA_in_Self*) (refer to Table OA.7 for detailed summary statistics).

We examine whether framing the task as weighting AI predictions versus original predictions induces different cognitive efforts during weight assignment by estimating the following model:[13]

[13] Analysis of PT for original predictions and feedback review showed no significant differences across conditions, with or without round number controls.

$$PT\_on\_weight_{it} = \beta\, condition_i + \varepsilon_{it} \tag{9}$$

In equation 9, the dependent variable $PT\_on\_weight_{it}$ represents the PT measured in seconds that participant i spent on weight assignment for AI or original predictions in round t. The independent variable is the experimental condition to which participant i was assigned. The results are presented in Table 10.

Table 10: Cognitive Effort on Weight Assignment

	PT_on_weight *Baseline: UA_in_AI*			
	(1)	(2)	(3)	(4)
No_AI	-2.35^{***}	-2.45^{***}	-2.26^{***}	-2.26^{***}
	(0.65)	(0.63)	(0.58)	(0.58)
BA_in_AI	1.95^{***}	1.88^{***}	1.99^{***}	2.00^{***}
	(0.73)	(0.71)	(0.67)	(0.67)
BA_in_Self	0.79	0.84	0.88	0.88
	(0.69)	(0.66)	(0.61)	(0.61)
PT_on_original		0.06^{**}	0.04^{**}	0.04^{**}
		(0.02)	(0.02)	(0.02)
lag_PT_on_feed			0.04	0.04
			(0.03)	(0.02)
Round				-0.22^{***}
				(0.02)
Constant	7.82^{***}	7.14^{***}	6.78^{***}	9.13^{***}
	(0.53)	(0.46)	(0.44)	(0.49)
N	7,199	7,199	6,839	6,839
Adjusted R^2	0.03	0.04	0.04	0.06

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Standard errors are in parentheses and clustered at an individual level.

The results in column 1 show that participants in the *BA_in_AI* condition spent significantly more PT on weight assignment (1.95, $p < 0.01$) compared to the baseline *UA_in_AI* condition. In contrast, participants in the *No_AI* condition spent significantly less PT (-2.35, $p < 0.01$) compared to the baseline *UA_in_AI* condition.[14]. In columns 2-4, we add control variables:

[14] For the *No_AI* condition, although participants did not actively assign weights, we included it as a reference to evaluate the reduction in cognitive load with AI assistance

$PT\_on\_original_t$ (processing time participant spent on making their original prediction in each round), $PT\_on\_feed_{t-1}$ (processing time participant spent on reviewing feedback from the previous round), as well as round t. The results remain consistent across all models.

Notably, the *BA_in_Self* condition shows no significant difference in weight assignment PT compared to the *UA_in_AI* baseline. The increased PT spent on weight assignment for AI predictions compared to original predictions suggests two implications. First, it indicates higher cognitive load (Boyaci et al. 2024), as evidenced by the *No_AI* condition where participants spent the least amount of time due to the absence of weight assignment. Second, it suggests that participants exerted more cognitive effort in calibrating weights when working with AI predictions than with their original predictions, indicating more careful consideration during AI weight assignment.

While both bounded autonomy approaches yield similar accuracy improvements, they operate through distinct mechanisms. *BA_in_Self* eliminates the anchoring bias of the initial weight, enabling participants to better incorporate AI assistance despite paying less cognitive effort when assigning weights to their original predictions. In contrast, *BA_in_AI* drives participants to invest more cognitive effort in deciding the weight assignment for AI predictions, as evidenced by longer decision times. These findings reveal that mathematically equivalent constraints can trigger different cognitive responses: a reduction in anchoring versus increased cognitive effort while achieving similar performance improvements.

In summary, we demonstrate that the bounded autonomy approach improves human-AI prediction accuracy compared to unbounded autonomy. Under unbounded autonomy, participants exhibited suboptimal weighting due to *initial weight inertia*, manifesting as anchoring and ceiling effects of the initial weight on subsequent weighting decisions, and evidenced by neglect of performance feedback. Bounded autonomy leads to improved performance through feedback-based weight adjustments and reduced *initial weight inertia*. While both bounded autonomy approaches - assigning weights to AI predictions versus original predictions - yield similar accuracy improvements, they operate through distinct mechanisms: *BA_in_Self* eliminates anchoring bias of the initial weights, while *BA_in_AI* increases cognitive effort when deciding the weight for AI predictions.

6.3. Managerial Insight

Our findings suggest that organizations can generate value in human-AI collaboration through informed designs of autonomy instead of providing full autonomy to humans in weighting AI advice. We show the effectiveness of one such design, which we call the bounded autonomy approach, where AI autonomy is based on employees' decision-making quality. Bounded

autonomy can be framed with weights applied to AI or human (original) predictions as the locus. While both framings of bounded autonomy approaches achieve similar accuracy gains, they suit different employee profiles. For employees prone to anchoring on past decisions (e.g., those exhibiting consistent weightings despite poor performance), constraining autonomy in weighting their original predictions helps overcome cognitive biases. In contrast, for employees making quick, potentially under-analyzed decisions, constraining AI weighting autonomy promotes more thoughtful consideration of AI recommendations. Importantly, while these approaches improve prediction accuracy, they do so by enhancing employees' ability to effectively leverage AI rather than developing their fundamental prediction skills.

7. Discussion and Conclusion

AI predictions are increasingly becoming ubiquitous as decision input in many management scenarios. While users increasingly have access to highly accurate low-cost AI predictions (Agrawal et al. 2022), organizations are yet to gain efficiently from AI in decision-making. Effective utilization of AI in practice remains either limited or suboptimal (e.g., Dietvorst et al. 2015, Balakrishnan et al. 2024). How people incorporate AI advice in predictive tasks therefore remains a critical question in management literature as well as practice. This challenge is particularly salient given a recent emphasis on human agency and people's documented preference for their own predictions over AI recommendations, even when AI demonstrates superior accuracy. These dynamics create a paradox: inaccurate users may make poor decisions despite having access to accurate AI systems.

Our study examines several fundamental questions about human-AI collaboration in general prediction tasks. How do people incorporate AI predictions in decision-making when given unbounded autonomy in weighting AI predictions? Does performance feedback influence prediction quality in human-AI collaboration? What happens when autonomy shifts from an unconditional right to a performance-based privilege? We posit and find empirical evidence that bounded autonomy—where decision-making freedom is contingent on performance—provides crucial signals that help users better incorporate AI recommendations. In our study, we focus specifically on how autonomy influences the weights people assign to AI predictions in sequential tasks, by directly observing the weights rather than using the commonly employed post-hoc calculation of advice-weighting in human-AI collaboration studies (e.g., Logg et al. 2019).

Through an online experiment, we demonstrate that under unbounded autonomy, people exhibit *initial weight inertia* when incorporating AI predictions, leading to suboptimal performance. Notably, participants remain anchored to their initial weight and treat this initial weight as a ceiling despite receiving repeated performance feedback over time. Further analysis

reveals that bounded autonomy significantly improves decision accuracy by overcoming this inertia—enabling movement away from the anchored initial weight and encouraging higher weights after the initial round. Our battery of tests shows that this improvement stems not from mechanical constraints, but from increased response to feedback. To better understand how bounded autonomy achieves these improvements, we examined its effects across different framings. While overall performance improvements occur under both framings, the underlying mechanisms differ: when participants were asked to assign weights to their original predictions, this completely eliminated the anchoring effect. In contrast, when participants were asked to assign weights to AI predictions, their cognitive effort increased in making decisions.

Our work extends human-AI collaboration literature by examining how autonomy can lead to *initial weight inertia* and how performance-based bounded autonomy can mitigate these effects. While previous research has examined algorithm aversion, appreciation, and cognitive processing, the role of managerial autonomy remains underexplored, despite its recognized importance in general decision-making. Our work integrates these streams of literature by demonstrating that autonomy significantly influences AI integration strategies, adding a crucial dimension to existing research in human-AI collaboration. These findings have important implications for users and business leaders implementing AI in predictive tasks.

More specifically, our research suggests that merely providing accurate AI tools and unbounded autonomy is insufficient for optimal performance. When unbounded autonomy is given to inaccurate users, it inherently leads to suboptimal outcomes due to weight inertia. These findings suggest the need for policies that balance autonomy with managerial predictive accuracy, such as enhanced training programs and calibrated autonomy levels for new or underperforming users. This requires taking a balanced view of managerial agency that recognizes both the value and potential limitations of decisional freedom.

Our findings have several broader consequences for both academia and practice. First, simply incorporating accurate AI tools may not lead to optimal outcomes—the persistence of *initial weight inertia* reveals that mere exposure to these tools without careful consideration of weighting strategies undermines AI's potential benefits. Second, we find that feedback alone proves ineffective without consequences for weight selection. This finding opens new research directions in feedback system design, aligning with Kluger and DeNisi (1996) observation that two-thirds of feedback interventions show no or negative impact on performance.

Third, our findings have particular relevance for settings where AI incorporation creates either a crutch effect or leads to deskilling. While AI improves overall performance, users' independent predictive abilities remain unchanged despite feedback and bounded autonomy. This suggests that bounded autonomy serves primarily as a mechanism for improving AI weight selection rather

than enhancing innate human performance. Instead of learning from feedback to become better predictors themselves, users develop better judgment about when and how much to rely on AI. This finding contributes to ongoing debates about AI's role in skill development and potential skill reduction across various tasks.

In conclusion, our work demonstrates that unrestricted autonomy leads to *initial weight inertia* and suboptimal AI incorporation. Organizations implementing AI systems should calibrate users' degree of autonomy based on their predictive accuracy. While performance-linked autonomy improves outcomes, it primarily enhances AI utilization rather than independent human prediction skills. These findings challenge the assumption of unconditional human agency in AI integration and reveal important nuances in human-AI collaboration, feedback effectiveness, and learning outcomes. Our work extends beyond human-AI collaboration to inform the broader understanding of feedback mechanisms and skill development in technology-augmented decision-making.

References

Agrawal A, Gans J, Goldfarb A (2022) *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence* (Harvard Business Press).

Balakrishnan M, Ferreira K, Tong J (2024) Human-algorithm collaboration with private information: Naïve advice weighting behavior and mitigation. *Available at SSRN 4298669* .

Bandura A (1991) Social cognitive theory of self-regulation. *Organizational behavior and human decision processes* 50(2):248–287.

Bastani H, Bastani O, Sinchaisri WP (2021) Learning best practices: Can machine learning improve human decision-making. *Academy of Management Proceedings*, volume 1, 14006 (Academy of Management Briarcliff Manor, NY 10510).

Beer R, Qi A, Rios I (2024) Behavioral externalities of process automation. *Available at SSRN 4295527* .

Bonaccio S, Dalal RS (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.

Boyaci T, Canyakmaz C, de Véricourt F (2024) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* 70(2):1258–1275.

Brau R, Aloysius J, Siemsen E (2023) Demand planning for the digital supply chain: How to integrate human judgment and predictive analytics. *Journal of operations management* 69(6):965–982.

Buell RW, Mariadassou S, Zheng Y (2019) Relative performance transparency: Effects on sustainable choices. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (19-079).

Burton JW, Stein MK, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making* 33(2):220–239.

Caro F, de Tejada Cuenca AS (2023) Believing in analytics: Managers' adherence to price recommendations from a dss. *Manufacturing & Service Operations Management* 25(2):524–542.

Chen C, Jain N, Karamshetty V (2023) Algorithm-human-algorithm: A new classification approach to integrating judgemental adjustments .

Chen DL, Schonger M, Wickens C (2016) otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97.

Choudhary V, Marchetti A, Shrestha YR, Puranam P (2025) Human-ai ensembles: When can they work? *Journal of Management* 51(2):536–569.

Choudhary V, Shunko M, Netessine S (2021) Does immediate feedback make you not try as hard? a study on automotive telematics. *Manufacturing & Service Operations Management* 23(4):835–853.

Dahlinger A, Wortmann F, Ryder B, Gahr B (2018) The impact of abstract vs. concrete feedback design on behavior insights from a large eco-driving field experiment. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–11.

Dai T, Abràmoff MD (2023) Incorporating artificial intelligence into healthcare workflows: Models and insights. *Tutorials in Operations Research: Advancing the Frontiers of OR/MS: From Methodologies to Applications*, 133–155 (INFORMS).

Dargnies MP, Hakimov R, Kübler D (2024) Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science* .

Dietvorst BJ, 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 BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144(1):114.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science* 64(3):1155–1170.

Epley N, Gilovich T (2006) The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological science* 17(4):311–318.

Fink L, Newman L, Haran U (2024) Let me decide: Increasing user autonomy increases recommendation acceptance. *Computers in Human Behavior* 156:108244.

Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research* 33(2):678–696.

Gino F, Moore DA (2007) Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making* 20(1):21–35.

Gnewuch U, Morana S, Heckmann C, Maedche A (2018) Designing conversational agents for energy feedback. *Designing for a Digital and Globalized World: 13th International Conference, DESRIST 2018, Chennai, India, June 3–6, 2018, Proceedings 13*, 18–33 (Springer).

Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19(1):121–127.

Green B (2022) The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45:105681.

Greene RL (1986) Sources of recency effects in free recall. *Psychological Bulletin* 99(2):221.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Ibrahim R, Kim SH (2019) Is expert input valuable? the case of predicting surgery duration. *Seoul Journal of Business, Forthcoming* .

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* 32(3):713–735.

Kawaguchi K (2021) When will workers follow an algorithm? a field experiment with a retail business. *Management Science* 67(3):1670–1695.

Kellogg KC, Valentine MA, Christin A (2020) Algorithms at work: The new contested terrain of control. *Academy of management annals* 14(1):366–410.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science* 66(11):5182–5190.

Kim SH, Song H (2022) How digital transformation can improve hospitals' operational decisions. *Harvard Business Review [Internet]* .

Kluger AN, DeNisi A (1996) The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin* 119(2):254.

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science* 52(1):111–127.

Lehmann CA, Haubitz CB, Fügener A, Thonemann UW (2022) The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Production and Operations Management* 31(9):3419–3434.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Luong A, Kumar N, Lang KR (2020) Algorithmic decision-making: examining the interplay of people, technology, and organizational practices through an economic experiment. *Baruch College Zicklin School of Business Research Paper* (2020-02):03.

Önkal D, Goodwin P, Thomson M, Gönül S, Pollock A (2009) The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22(4):390–409.

Paas FG, Van Merriënboer JJ (1994) Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review* 6:351–371.

Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: An attentional integration. *Human factors* 52(3):381–410.

Raisch S, Krakowski S (2021) Artificial intelligence and management: The automation–augmentation paradox. *Academy of management review* 46(1):192–210.

Samuelson W, Zeckhauser R (1988) Status quo bias in decision making. *Journal of risk and uncertainty* 1:7–59.

Sarter NB, Schroeder B (2001) Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human factors* 43(4):573–583.

Snyder C, Keppler S, Leider S (2024) Algorithm reliance, fast and slow. *Fast and Slow (May 31, 2024)* .

Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of experimental psychology: Learning, memory, and cognition* 35(3):780.

Soule D, Grushka-Cockayne Y, Merrick J (2023) A heuristic for combining correlated experts when there are few data. *Management Science* .

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2022) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science* 68(2):846–865.

Sweller J (1988) Cognitive load during problem solving: Effects on learning. *Cognitive science* 12(2):257–285.

Tiefenbeck V, Goette L, Degen K, Tasic V, Fleisch E, Lalive R, Staake T (2018) Overcoming salience bias: How real-time feedback fosters resource conservation. *Management science* 64(3):1458–1476.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185(4157):1124–1131.

Yaniv I, Kleinberger E (2000) Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes* 83(2):260–281.

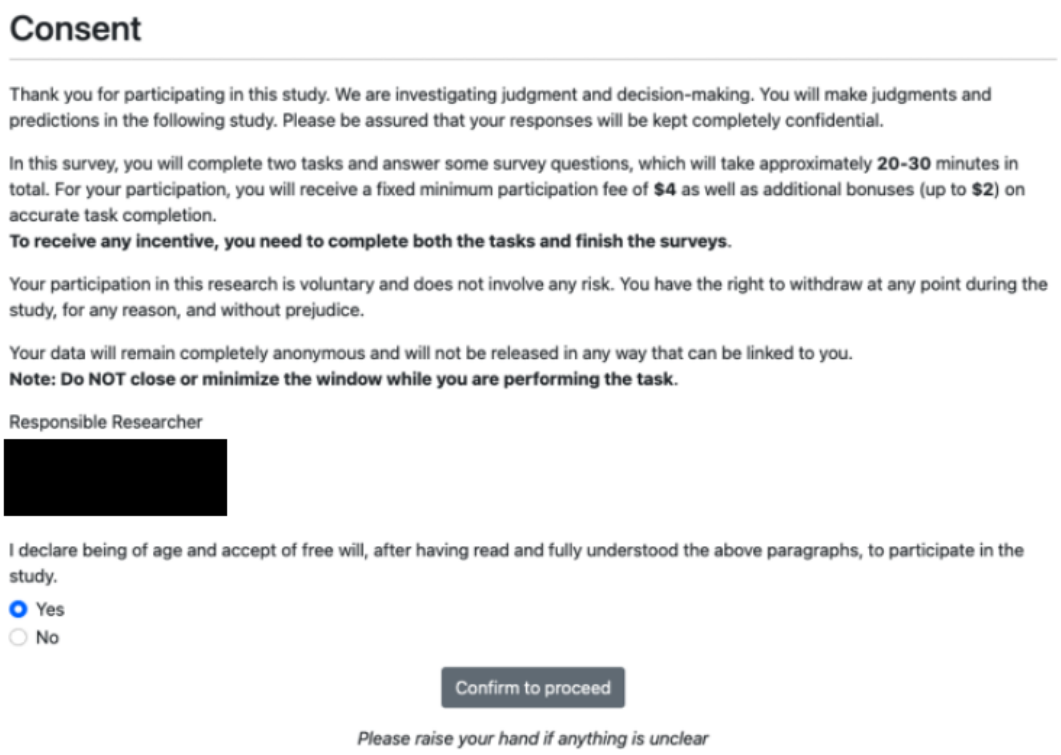
You S, Yang CL, Li X (2022) Algorithmic versus human advice: does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems* 39(2):336–365.

Online Appendix: *Does Autonomy Make People Try Less Hard? Initial Weight Inertia in Human-AI Collaboration*

1. Online Appendix A: Experiment Instructions

The experiment was computer-based, implemented using oTree (Chen et al. 2016), and conducted online via Prolific. This appendix presents the structure and instructions of the experiment. Participants began the experiment with the consent shown in Figure OA.1.

Figure OA.1: Consent of the Experiment



Note: A black box has been applied to the figure to comply with the blindness requirement of the review process.

Then, participants were asked about their gender and education level, as shown in Figure OA.2, to start the experiment.

Step 1. General Instructions

Participants received general instructions about the experiment's context, including key definitions of absolute errors and the bonus scheme based on the *MAE*. To ensure understanding, participants answered two comprehension questions. They must answer both correctly to proceed, as shown in Figure OA.3.

Figure OA.2: Gender and Education Level

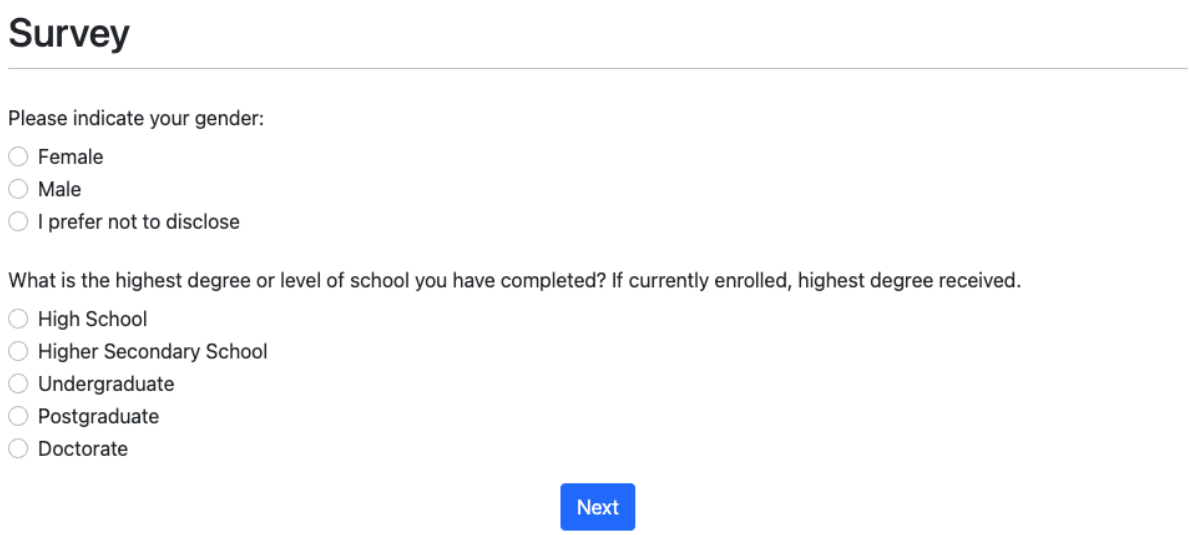
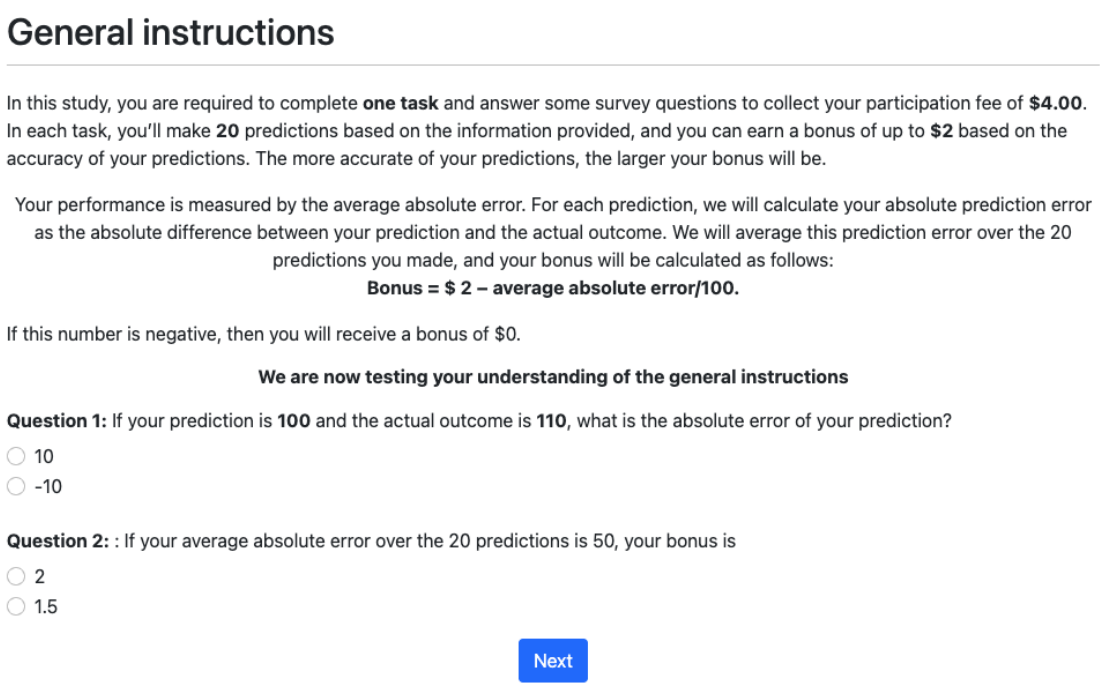


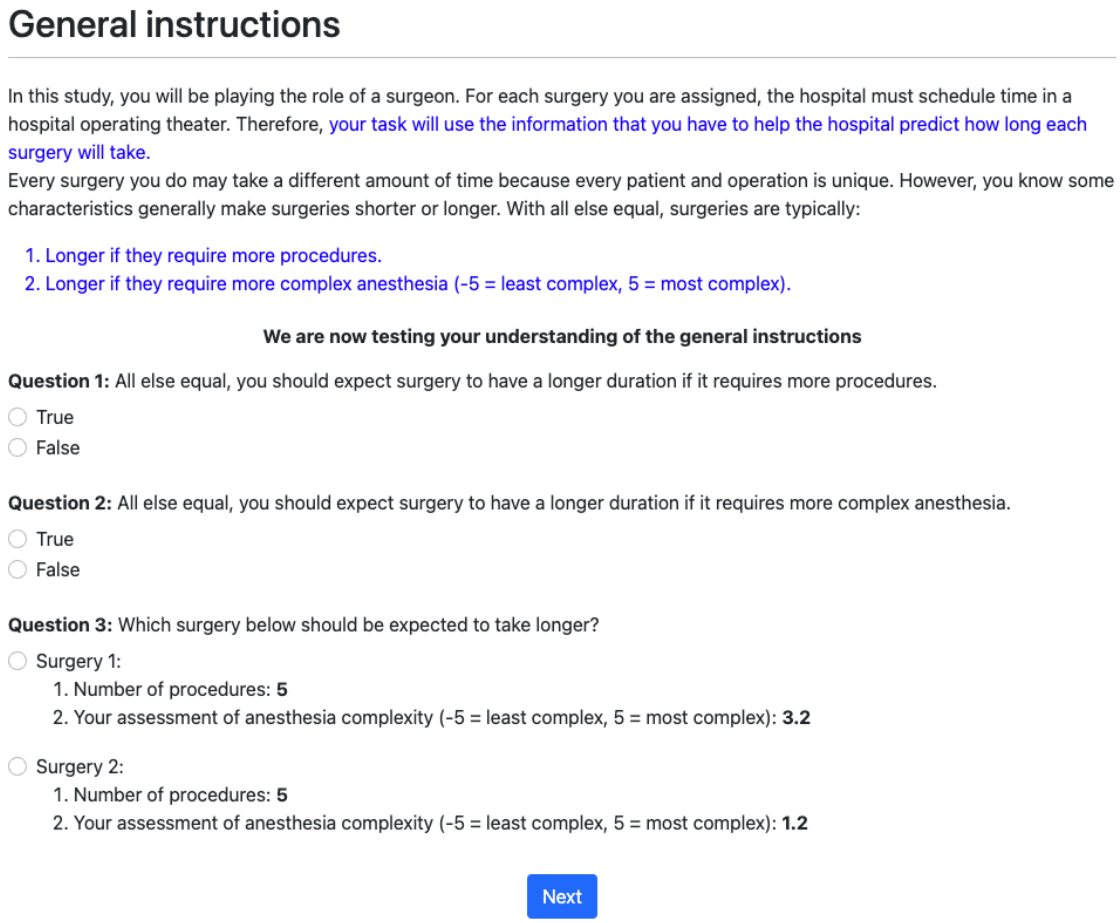
Figure OA.3: General instruction



Step 2. Historical Surgery Review

Participants were provided with instructions on how to make predictions based on two variables that determine the duration of surgery: the procedure and the complexity of the anesthesia. They were guided that longer procedures and more complex anesthesia would result in longer surgery durations. Participants were required to pass a three-question comprehension test before proceeding to the next step, as shown in Figure OA.4.

Figure OA.4: General instruction for the predictive variables



Then, participants were presented with 10 rounds of the historical surgery duration with two variables. They can attempt to guess the duration first before saying the actual outcomes, shown in OA.5.

After the review, participants were asked to rate how much they think about the difficulty of the tasks, shown in Figure OA.6.

Figure OA.5: Historical Surgery Review

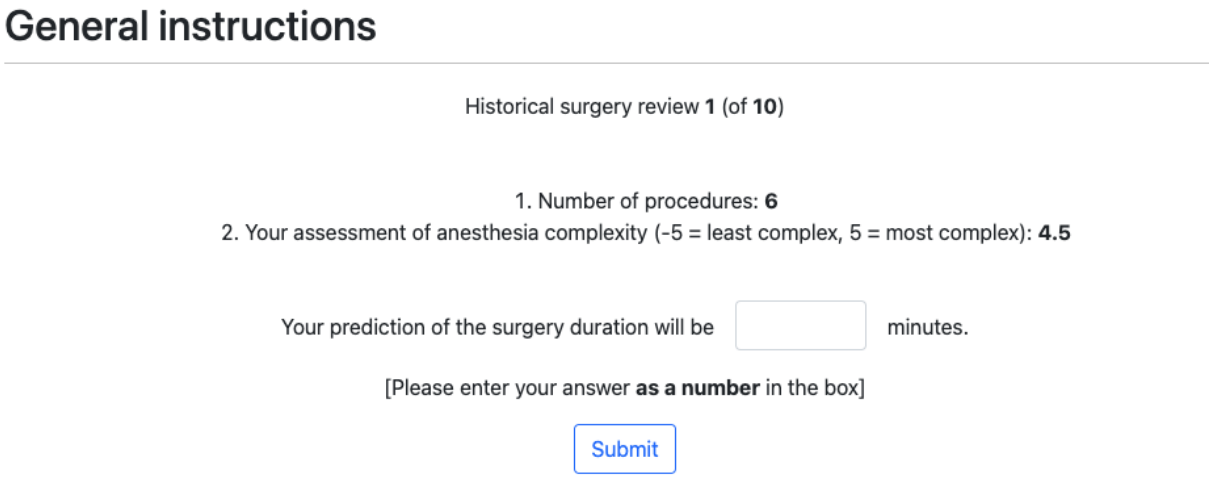
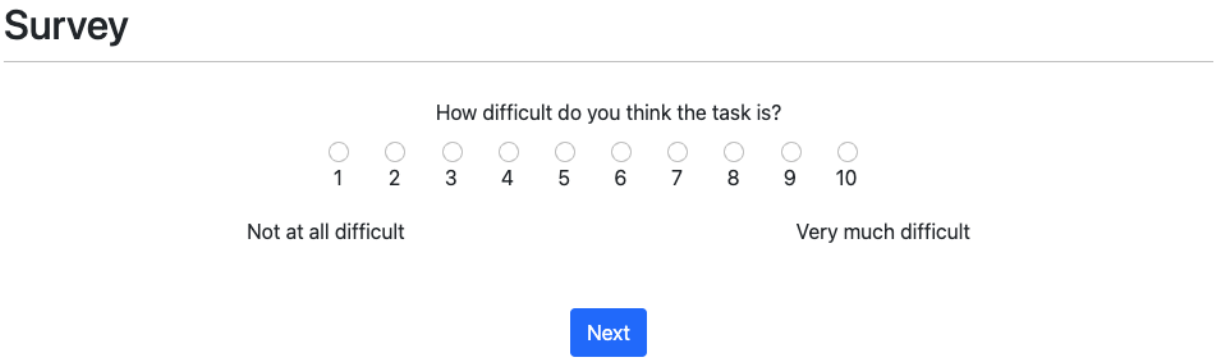


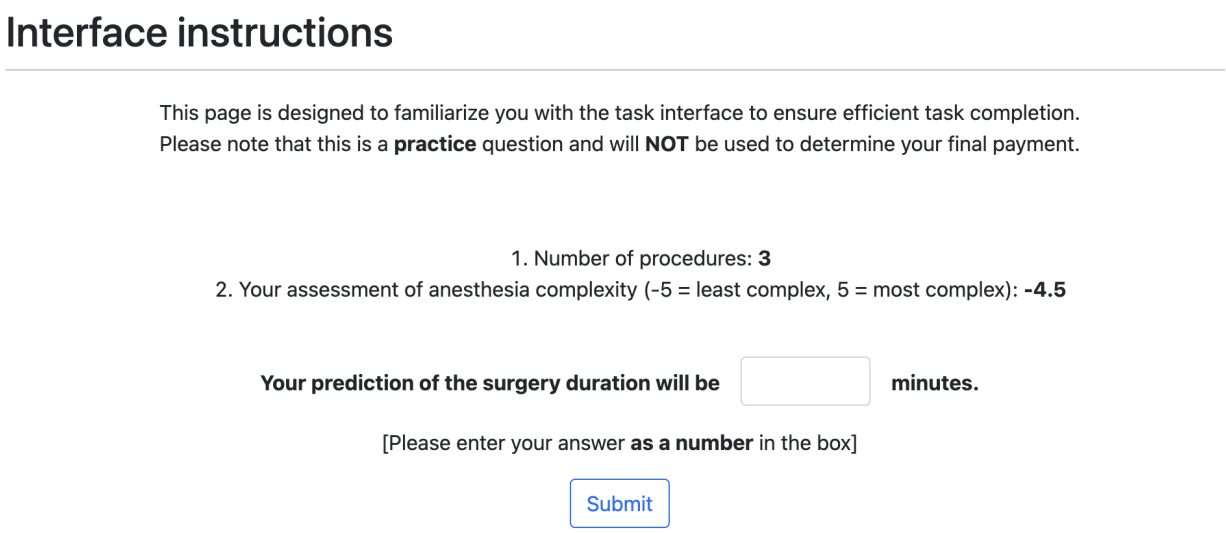
Figure OA.6: Perceived practice Difficulty



Step 3. Prediction Task

Participants were prompted to forecast the durations of 20 new surgeries not previously seen during training for the actual prediction task. Before making predictions, the interface instructions corresponding to the assigned experimental condition were displayed to each participant. The procedure consisted of two prediction stages. First, all participants generated an original prediction without assistance, facing the same interface instruction as Figure OA.7:

Figure OA.7: Interface Instruction Seen by All Participants



Next, they had the opportunity to revise this initial forecast with or without the algorithm's prediction, depending on the assigned experimental condition. The user interface instructions varied across the four conditions in the revision stage as follows from Figure OA.8 to Figure OA.17:

Then, participants moved to the 20 rounds of actual prediction tasks under their experimental conditions, as shown from Figure OA.18 to Figure OA.22.

After each round of predictions, participants were informed of the actual surgery duration, their final revised prediction, and the absolute prediction error, which is the absolute difference between the actual duration and the participant's revised prediction (Figure OA.23).

Figure OA.8: Revision Interface Instruction Seen by *No_AI* Condition

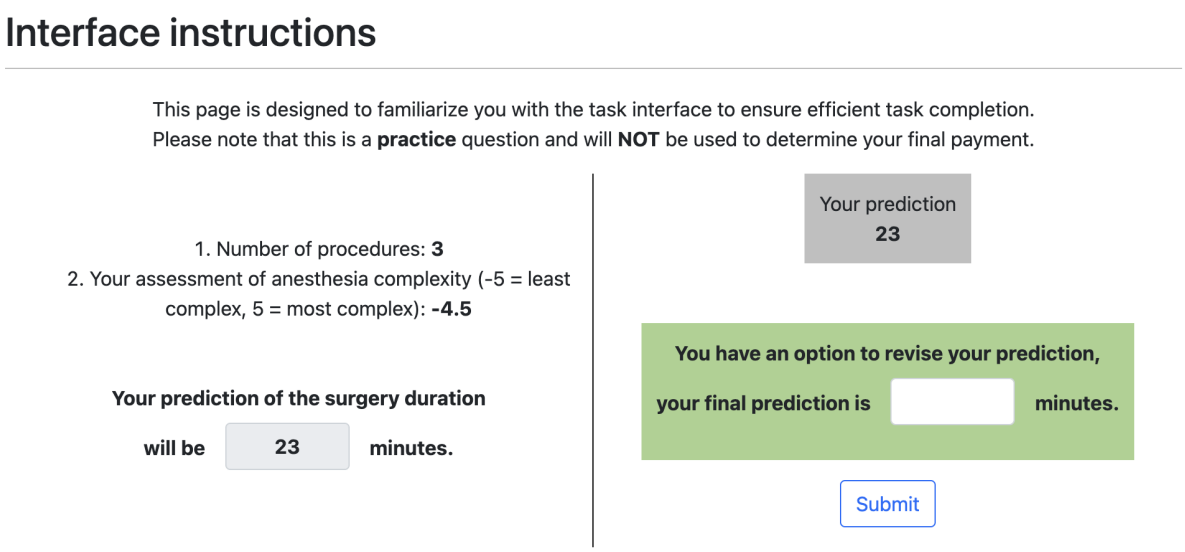


Figure OA.9: Revision Interface Instruction Seen by *UA_in_AI* Condition

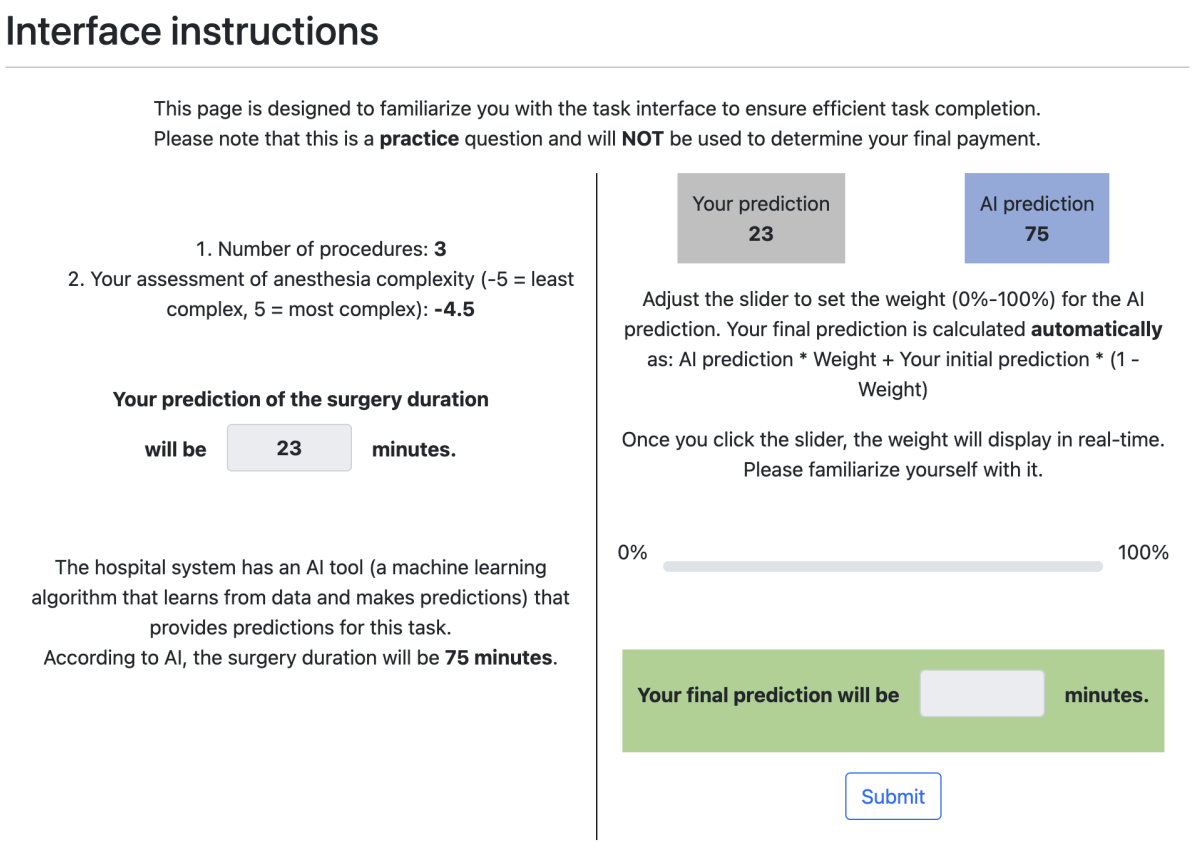


Figure OA.10: Revision Interface Instruction Seen by *BA_in_AI* Condition

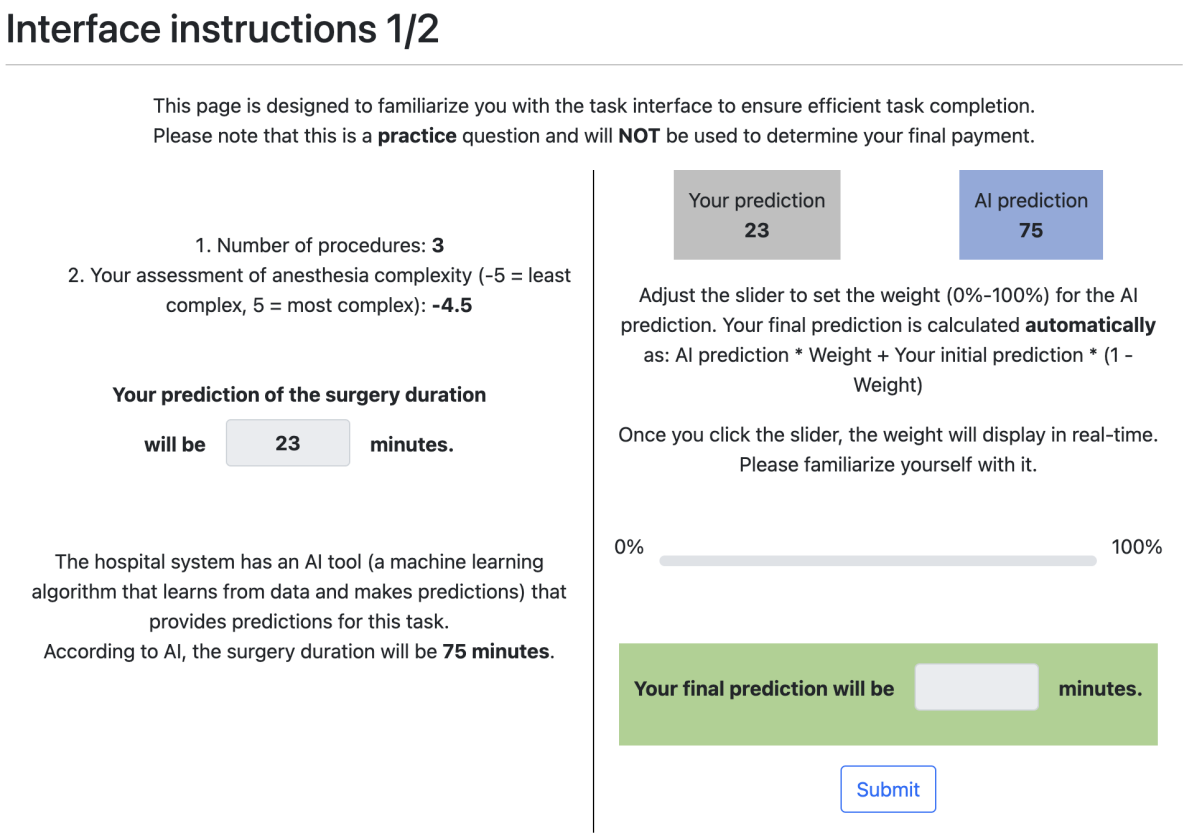


Figure OA.11: Revision Interface Seen by *BA_in_AI* Condition

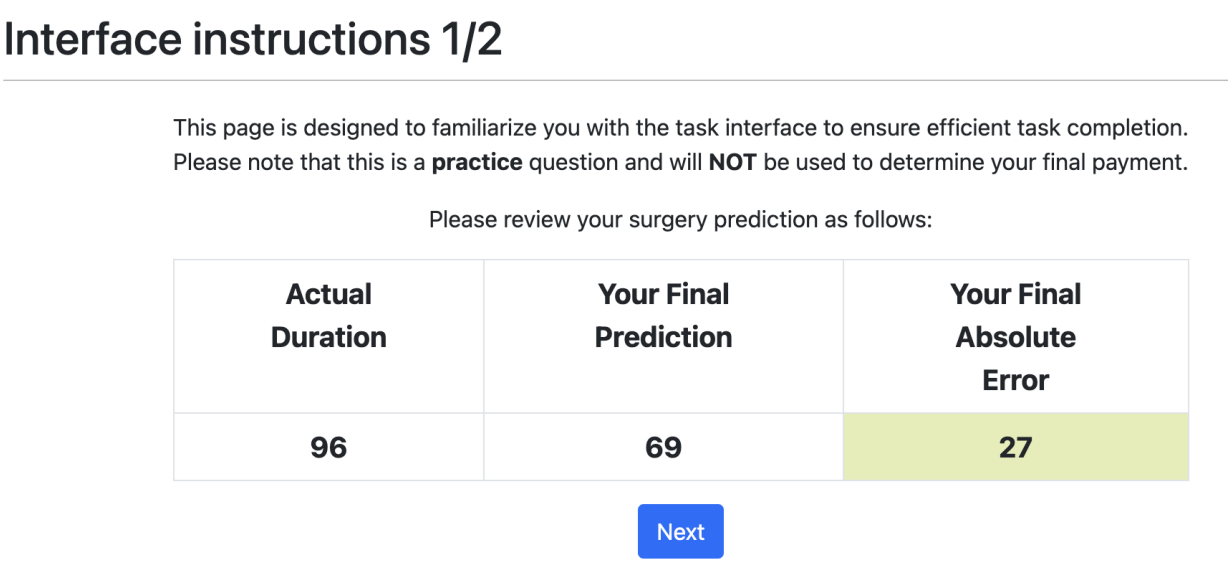


Figure OA.12: Revision Interface Seen by *BA_in_AI* Condition

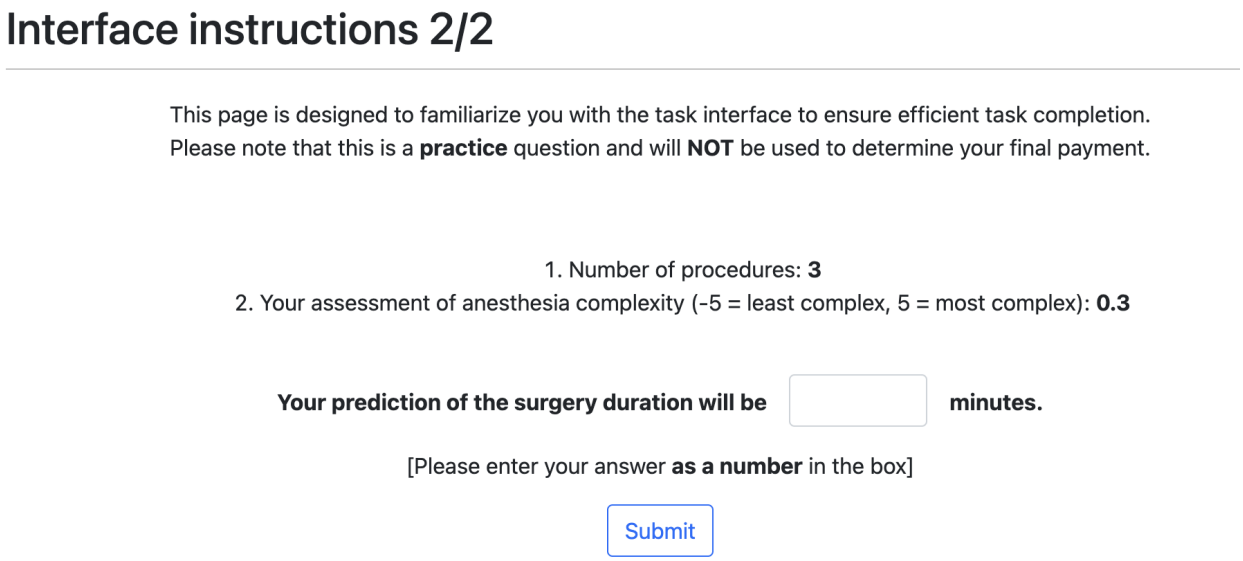


Figure OA.13: Revision Interface Seen by *BA_in_AI* Condition

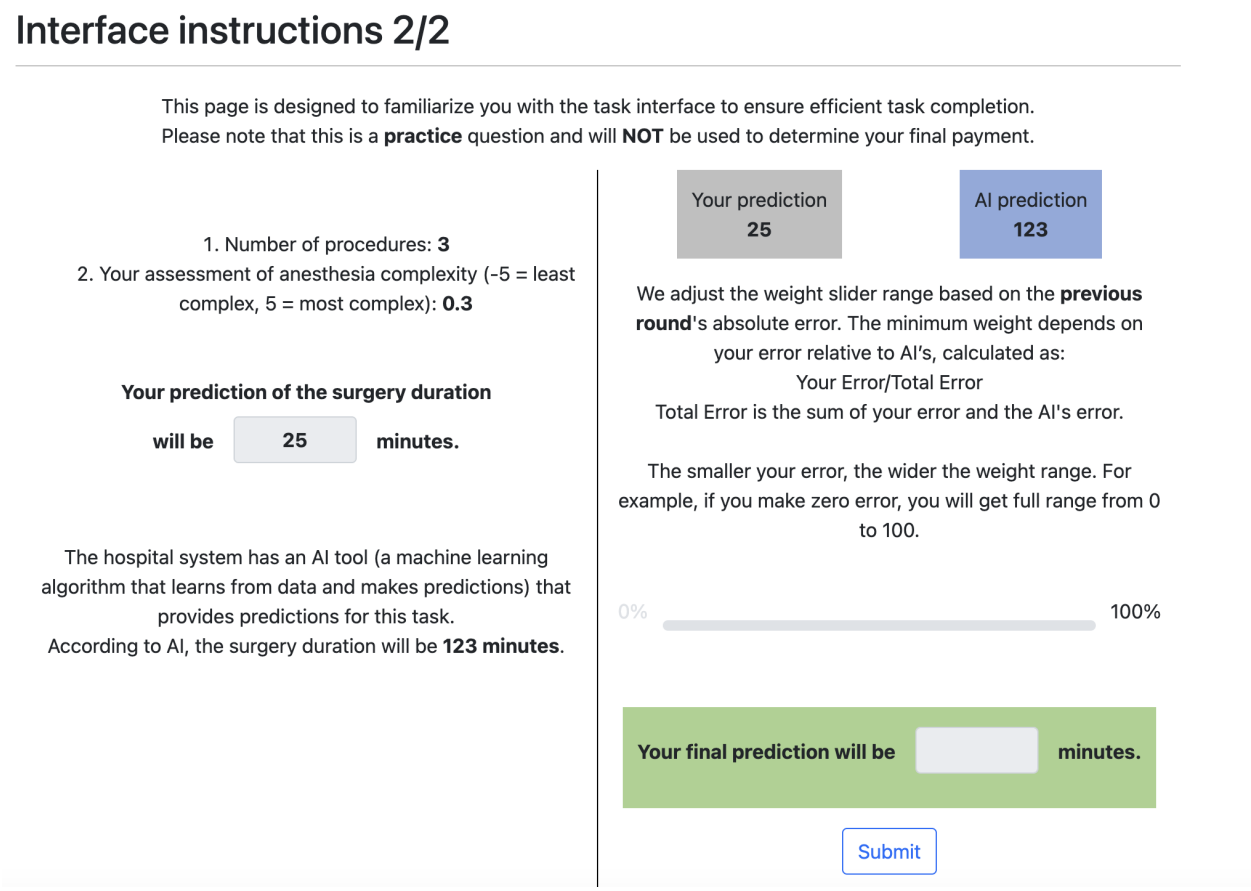


Figure OA.14: Revision Interface Instruction Seen by *BA_in_Self* Condition

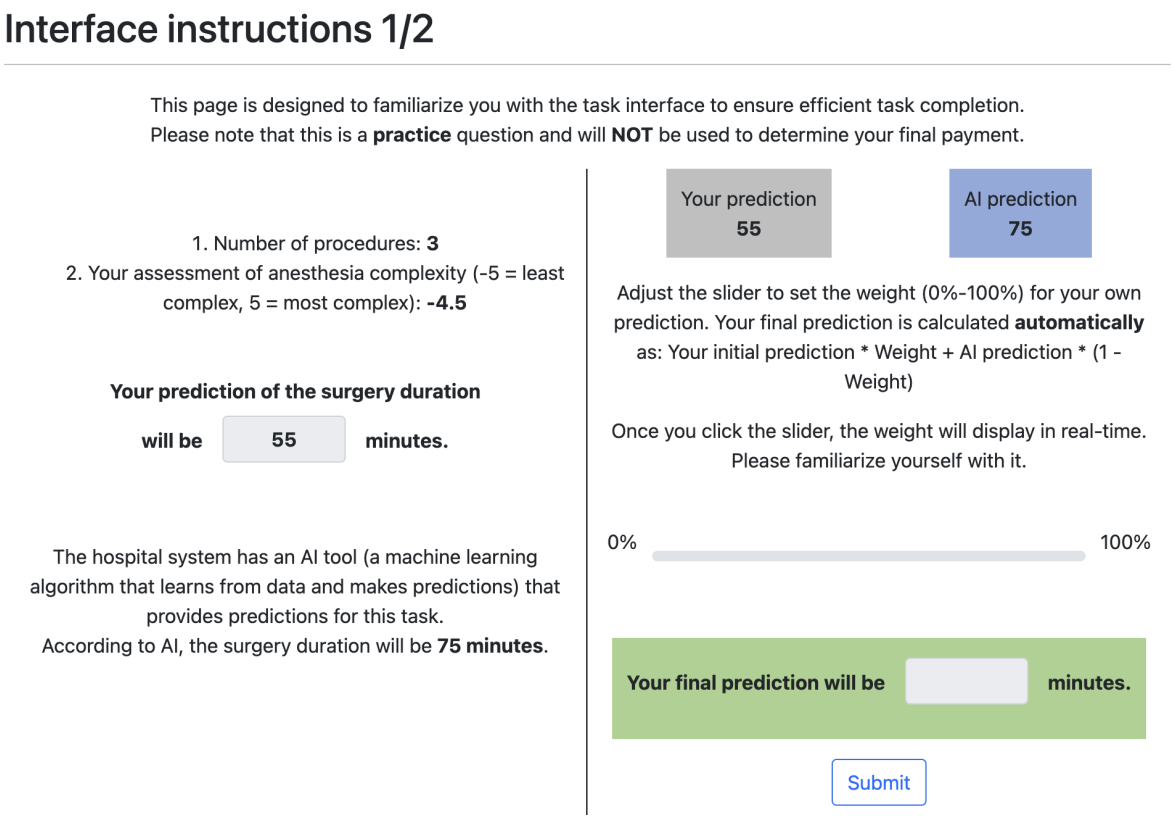


Figure OA.15: Revision Interface Instruction Seen by *BA_in_Self* Condition

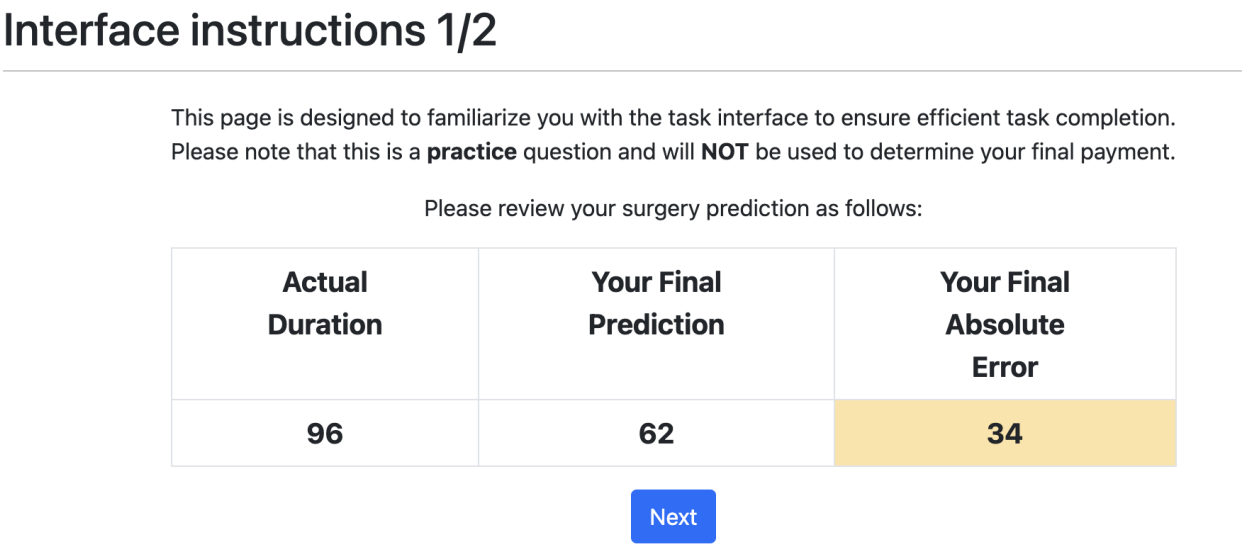


Figure OA.16: Revision Interface Instruction Seen by *BA_in_Self* Condition

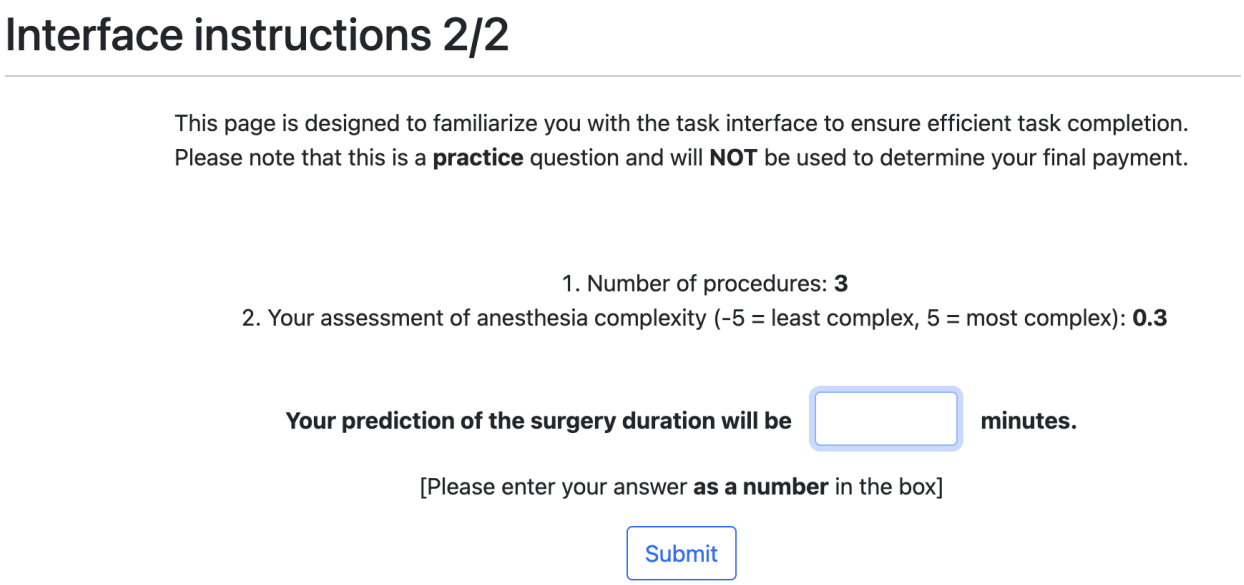


Figure OA.17: Revision Interface Instruction Seen by *BA_in_Self* Condition

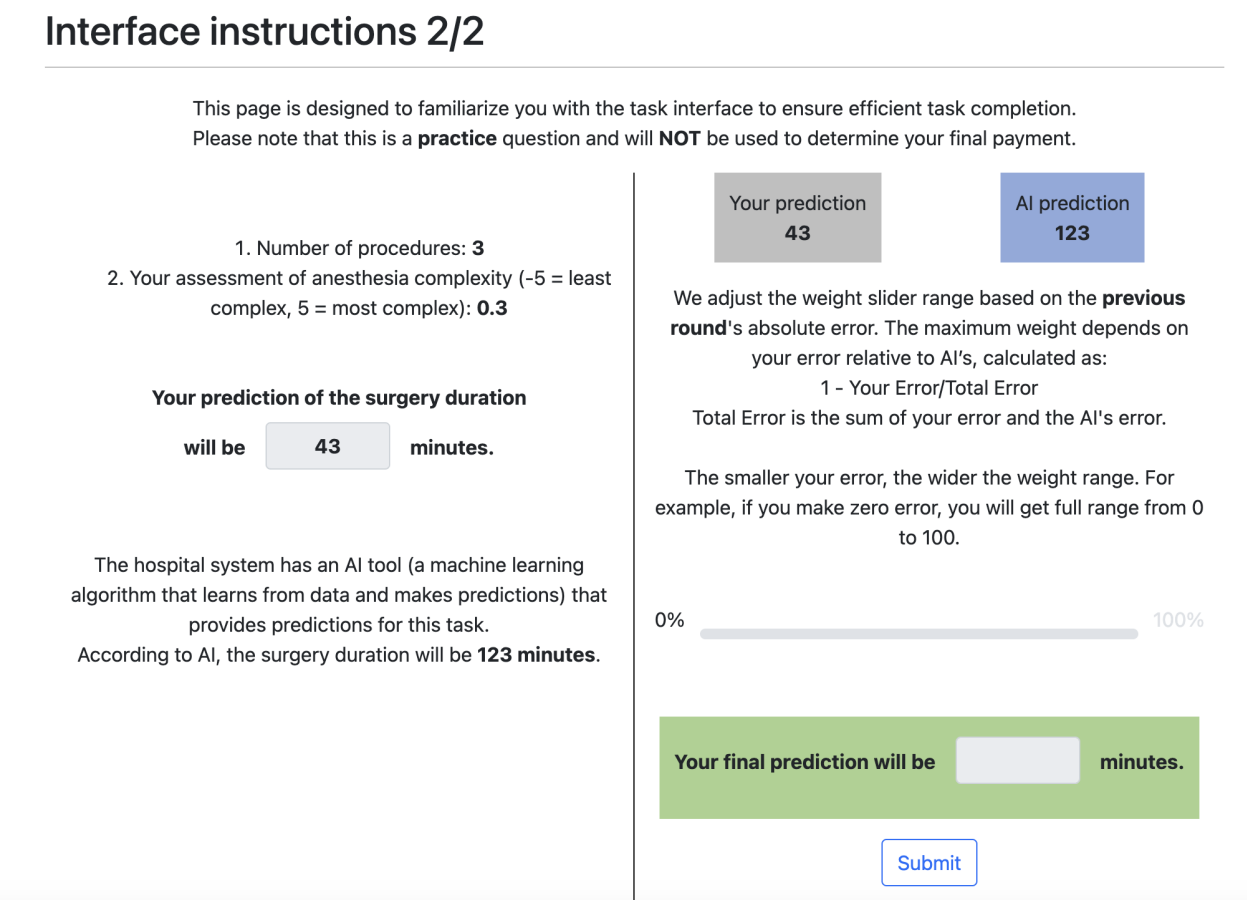


Figure OA.18: Initial Prediction Interface Seen by All Participants

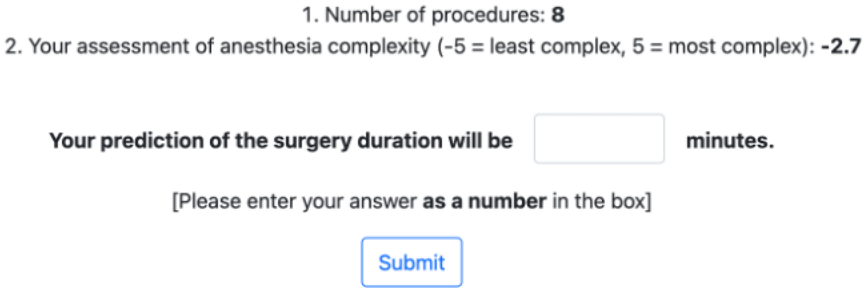


Figure OA.19: Revision Interface Seen by *No_AI* Condition

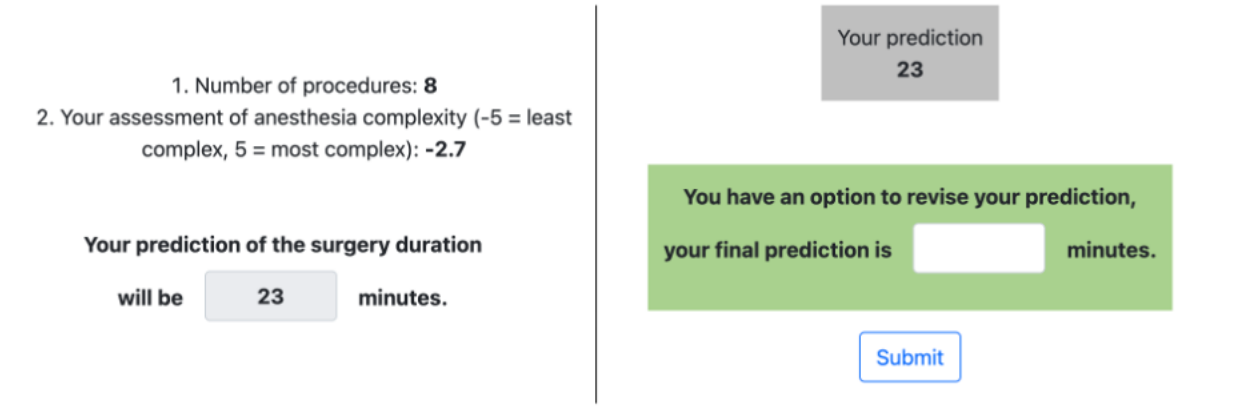
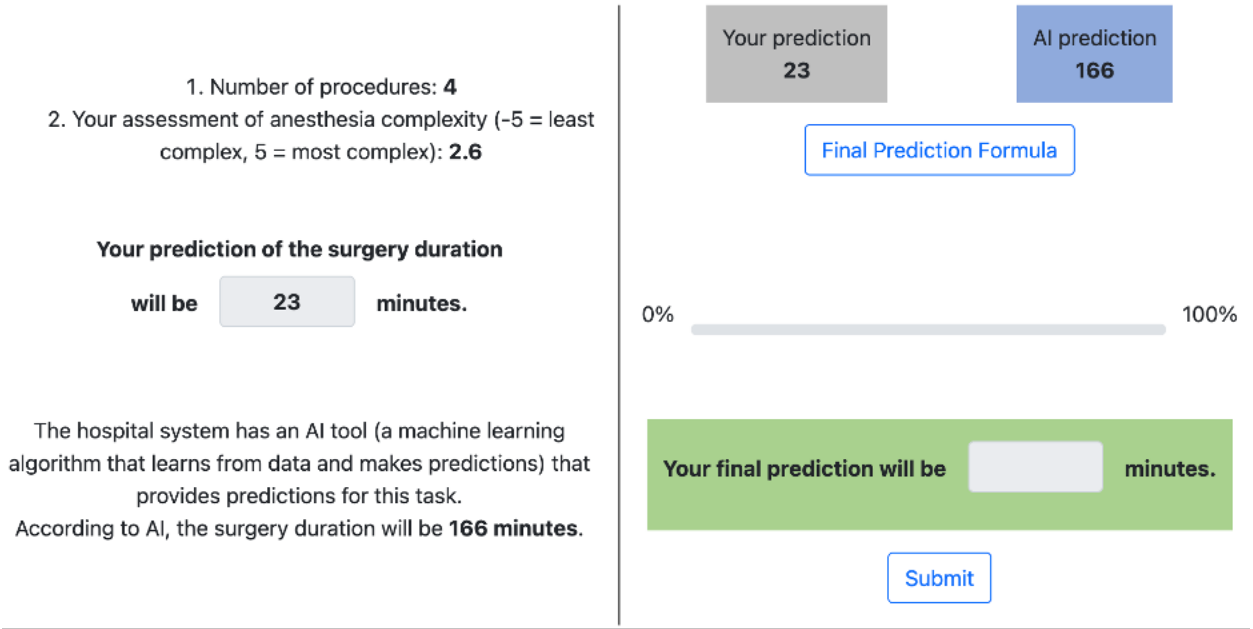
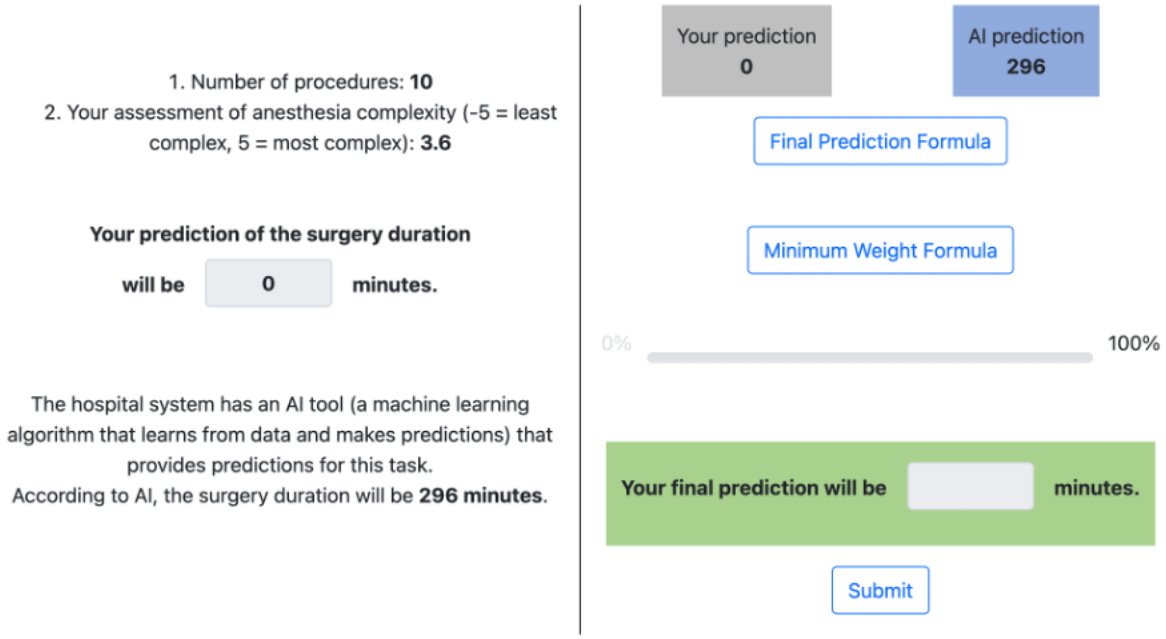


Figure OA.20: Revision Interface Seen by *UA_in_AI* Condition



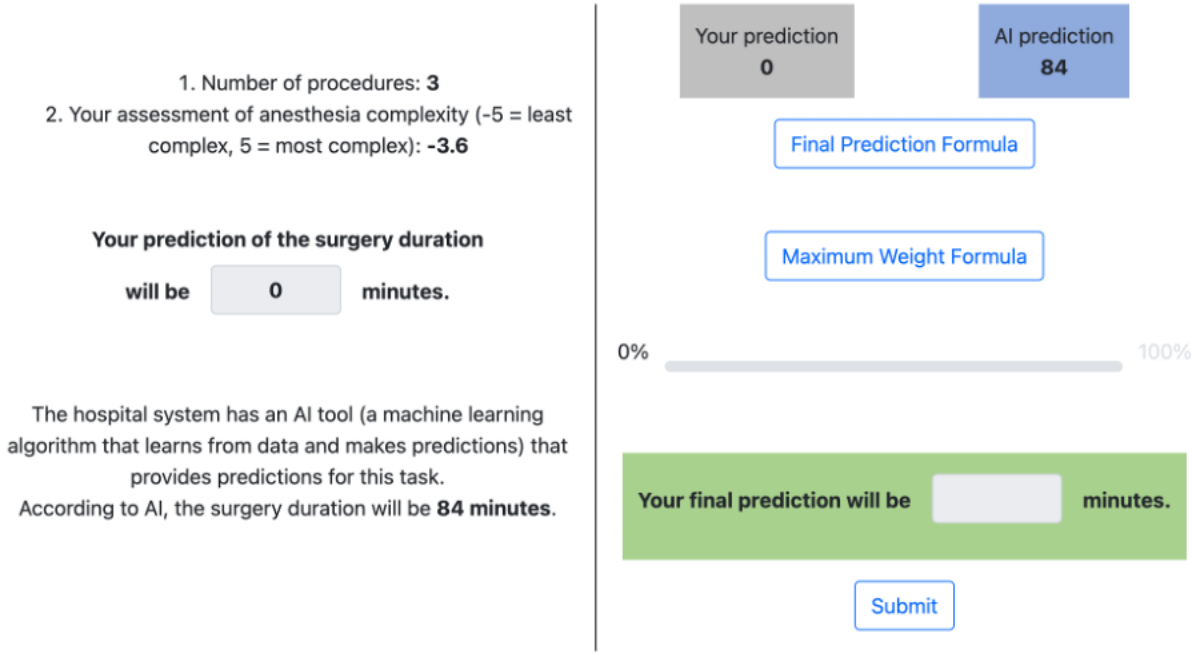
Note: Weight range was from 0% to 100% across the 20 prediction rounds.

Figure OA.21: Interface Seen by *BA_in_AI* Condition



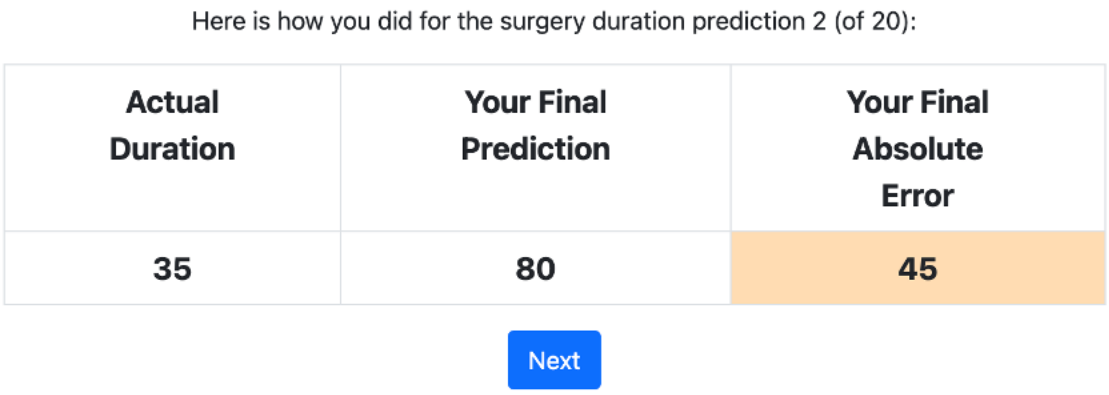
Note: The allowable weight range was dynamic, as illustrated by the interface conditions, with a defined minimum weight. Weights below the minimum were disabled and are shaded gray.

Figure OA.22: Revision Interface Seen by *BA_in_Self* Condition



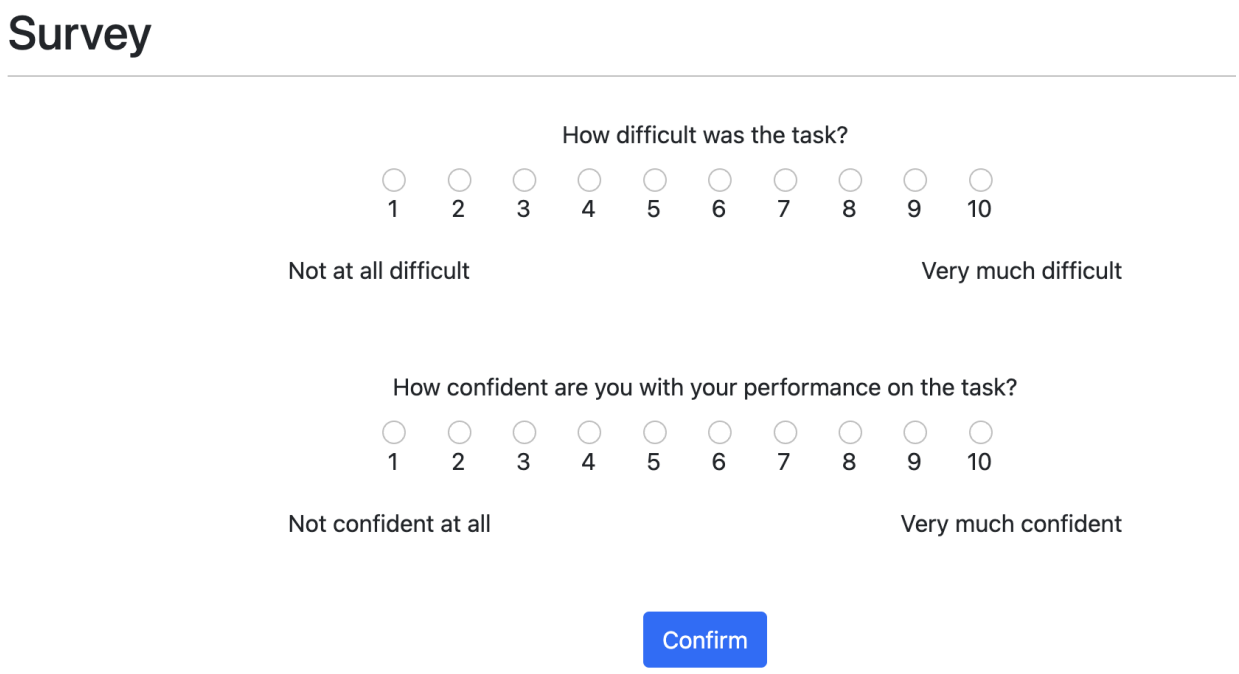
Note: The allowable weight range was dynamic, as illustrated by the interface conditions, with a defined maximum weight. Weights below the maximum were disabled and are shaded gray.

Figure OA.23: **Performance Feedback Interface Seen by All Participants.**



Step 4. Surveys After completing 20 rounds of prediction tasks, all participants were surveyed with questions to assess the perceived difficulty of the prediction task and their confidence in their predictions, shown in Figure OA.24.

Figure OA.24: Survey for All Participants



For those with AI assistance, participants were required to answer survey questions regarding AI, as shown in Figure OA.25.

Step 5. Payment Finally, participants were informed of their payment, including the bonus based on their prediction performance, and provided with a link to return to Prolific, as shown in Figure OA.26.

Figure OA.25: Survey for Participants with AI Assistance

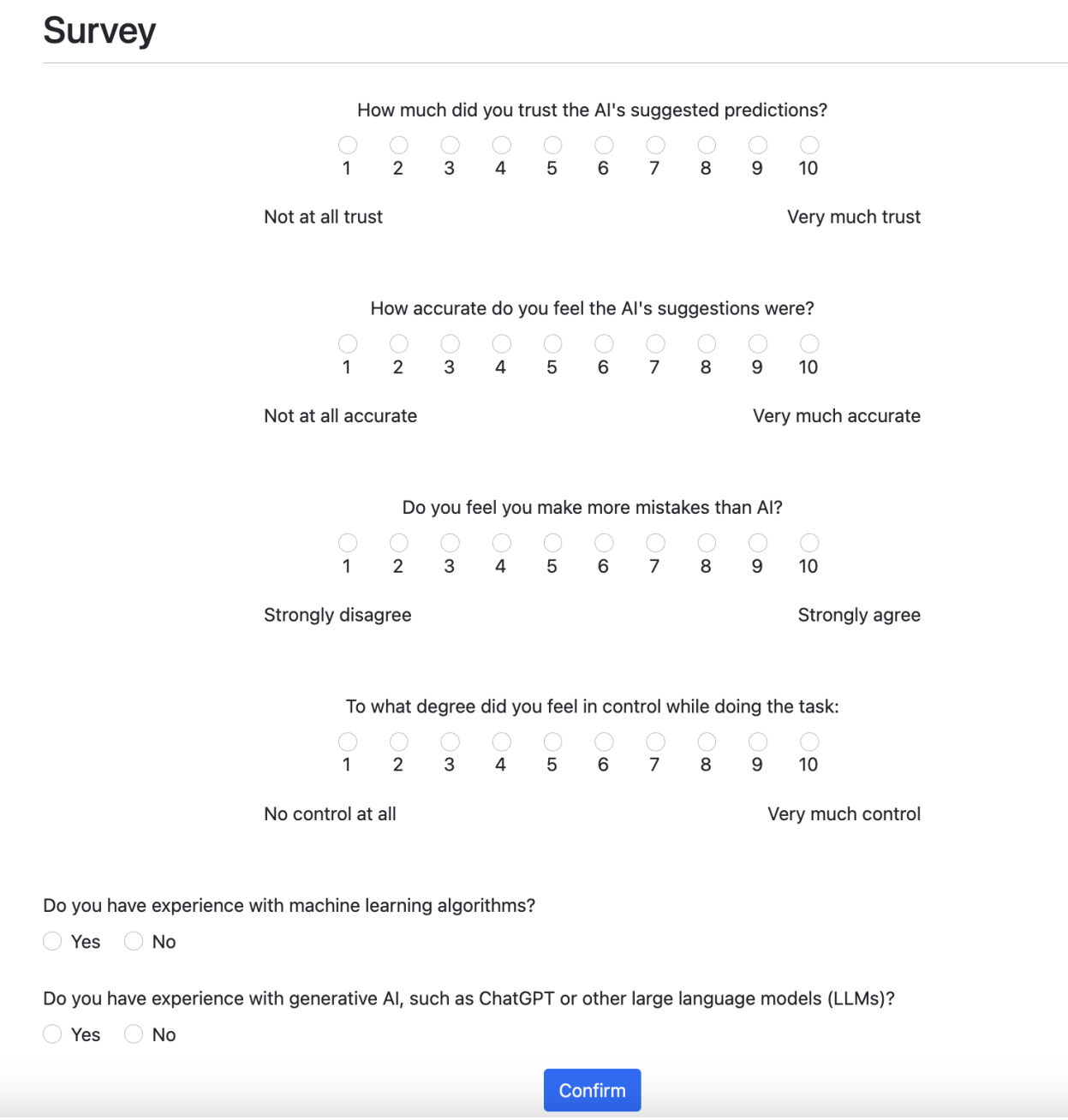
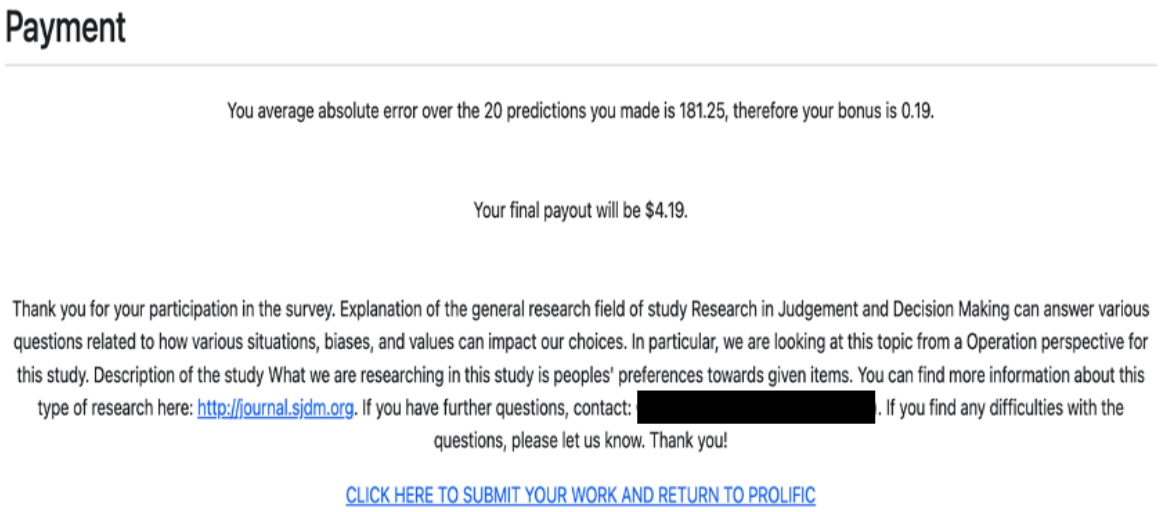


Figure OA.26: Payment



Note: A black box has been applied to the figure to comply with the blindness requirement of the review process.

2. Online Appendix B: Optimal Weight

With the assistance of the AI algorithm, participants make their revised predictions by combining their original predictions with the AI's predictions. Participants allocate WOA_t and to their original prediction to compute the revised $\hat{Y}_t^{\text{revised}}$ as follows:

$$\hat{y}_{it}^{\text{revised}} = \hat{y}_{it}^{\text{original}} \times (1 - WOA_{it}) + \hat{y}_{it}^{\text{AI}} \times WOA_t \tag{10}$$

To make the final revised prediction the same as the actual outcome y_i, we can calculate the optimal weight the participants can put in their weight on AI and their original prediction as follows:

$$WOA_{it}^{\text{optimal}} = \frac{y_{it} - \hat{y}_{it}^{\text{original}}}{\hat{y}_{it}^{\text{AI}} - \hat{y}_{it}^{\text{original}}} \tag{11}$$

Here, $WOA_{it}^{\text{optimal}}$ denotes the optimal weight that the participant should place on the AI prediction $\hat{y}_{it}^{\text{AI}}$ when the actual outcome is y_{it}.

Thus, for the unbounded autonomy in weighting AI prediction, where the weight WOA_{it} can range from 0% to 100% (i.e., $0 \leq WOA_{it} \leq 1$), we consider the following adjustments:

- If the optimal weight $WOA_{it}^{\text{optimal}} < 0$:
  - —Participants should set $WOA_{it} = 0$, placing no weight on the AI prediction.
- If the optimal weight $WOA_{it}^{\text{optimal}} > 1$:
  - —Participants should set $WOA_{it} = 1$, placing full weight on the AI prediction.

For the bounded autonomy in weighting AI prediction, where the weight WOA_{it} can range from min% to 100% (i.e., $WOA_{it}^{min} \leq WOA_{it} \leq 1$), where the WOA_{it}^{min} is adjusted each round based on the Equation 4 for each round t, we consider the following adjustments:

- If the optimal weight $WOA_{it}^{\text{optimal}} < WOA_{it}^{min}$:
  - —Participants should set the weight to $WOA_{it} = WOA_{it}^{min}$ on the AI prediction.
- If the optimal weight $WOA_{it}^{\text{optimal}} > 1$:
  - —Participants should set $WOA_{it} = 1$, placing full weight on the AI prediction.

For the bounded autonomy in weighting original predictions, where the weight WOS_{it} can range from 0% to max% (i.e., $0 \leq WOS_{it} \leq WOS_{it}^{max}$). The WOS_{it}^{max} is adjusted for each round based on the equation 5 for each round t, and mathematically WOA_{it} = 1 - WOS_{it}, therefore, we consider the following adjustments for WOS_{it} and WOA_{it}:

- If the optimal weight $WOS_{it}^{\text{optimal}} > WOS_{it}^{max}$:
  - —Participants should set the weight to $WOS_{it} = WOS_{it}^{max}$ on their original predictions, thus set the 1 - WOS_{it}^{max} on AI predictions.
- If the optimal weight $WOS_{it}^{\text{optimal}} < 0$:

— Participants should set $WOS_{it} = 0$, placing no weight on their original predictions, thus the 100% on AI predictions.

These adjustments ensure that the weights remain within the permissible range and reflect the optimal weighting decisions under unbounded and bounded autonomy.

3. Online Appendix C: Supplementary Analyses

Demographic Information Table OA.1 presents summary statistics of participants' demographic information (gender, education). Participants rated their trust in the AI, perceived AI accuracy, and perceived mistakes of their original predictions compared to the AI's, each on a scale from 0 to 10. They also rated their sense of control over the prediction process on the same scale. Additionally, we surveyed participants' prior experience with machine learning algorithms and generative AI, such as large language models like ChatGPT.

Table OA.1: Descriptive Statistics

Variable	**Description**	***No_AI***	***UA_in_AI***	***BA_in_AI***	***BA_in_Self***
Gender	Female = 0, Male = 1	0.49 (0.52)	0.51 (0.52)	0.54 (0.54)	0.51 (0.50)
Education	High school = 1, Higher Secondary School = 2, Undergraduate = 3, Postgraduate = 4, Doctorate = 5	2.44 (1.01)	2.47 (1.11)	2.72 (1.04)	2.69 (1.10)
Difficulty	How difficult was the task? (Scale from 1 to 10)	8.04 (1.98)	7.03 (1.82)	7.55 (1.89)	7.89 (2.14)
Confidence	How confident are you in your performance on the task? (Scale from 1 to 10)	4.22 (2.49)	5.00 (2.38)	4.80 (2.12)	4.63 (2.28)
Trust	How much did you trust the AI's suggested predictions? (Scale from 1 to 10)	–	6.65 (1.89)	6.43 (1.57)	6.16 (2.02)
Accuracy	How accurate do you feel the AI's suggestions were? (Scale from 1 to 10)	–	6.13 (1.84)	6.03 (1.65)	5.91 (1.85)
Mistakes	Do you feel you make more mistakes than AI? (Scale from 1 to 10)	–	7.21 (1.92)	6.78 (2.17)	6.90 (2.12)
Control	To what degree do you feel in control while doing the task? (Scale from 1 to 10)	–	6.81 (2.43)	5.83 (2.24)	5.53 (2.57)
Exp_ML	Do you have experience with machine learning algorithms? (Yes = 1, No = 0)	–	0.26 (0.44)	0.29 (0.46)	0.20 (0.40)
Exp_GAI	Do you have experience with generative AI, such as ChatGPT or other LLMs? (Yes = 1, No = 0)	–	0.71 (0.46)	0.81 (0.40)	0.79 (0.41)
N	Number of Participants	91	89	89	91

Practice Table OA.2 presents summary statistics of participants' MAE during the practice rounds and their perceived difficulty of the practice tasks. Participants did not exhibit any

significant differences in practice performance among the four experimental conditions, as measured by AE and perceived difficulty.

Table OA.2: Descriptive Statistics of Practice

Variable	Description	Treatment	Min	Mean (SD)	Max	N
AE_practice	The prediction errors, measured by AE, made by participants in the practice	*No_AI*	0	74.88 (337.83)	9747	910
		UA_in_AI	0	61.49 (69.02)	1338	890
		BA_in_AI	0	61.94 (60.89)	956	890
		BA_in_Self	0	62.54 (148.08)	3318	910
Difficulty_practice	How difficult was the task? (Scale from 1 to 10)	*No_AI*	2	7.80 (1.84)	10	910
		UA_in_AI	1	7.11 (2.17)	10	890
		BA_in_AI	2	7.63 (1.93)	10	890
		BA_in_Self	1	7.78 (1.94)	10	910

Post-Hoc Minimum and Maximum AE for Each AI-assisted Condition Table OA.3 presents summary statistics for the post-hoc calculation of minimum AE (best accuracy) and maximum AE (worst accuracy) under each condition.

Table OA.3: Summary Statistics of Minimum and Maximum AE for Each AI-Assisted Condition

Minimum AE	***UA_in_AI***	***BA_in_AI***	***BA_in_Self***
Min	0	0	0
Mean	19.78	23.04	23.00
SD	(30.54)	(31.18)	(30.78)
Max	100	100	100
N	1,780	1,780	1,820
Maximum AE	***UA_in_AI***	***BA_in_AI***	***BA_in_Self***
Min	0	1	0
Mean	62.36	43.37	41.93
SD	(49.14)	(38.79)	(35.85)
Max	362	471	240.4
N	1,780	1,780	1,820

Deviation from Optimal Weight Table OA.4 presents the summary statistics of the deviation from optimal weight across the three AI-assisted conditions.

Table OA.5 presents the summary statistics of the deviation from optimal weight across the three AI-assisted conditions.

WOA* Response to Performance Feedback Under *BA_in_Self

Table OA.4: Summary Statistics of Optimal Weight

Optimal Weight	***UA_in_AI***	***BA_in_AI***	***BA_in_Self***
Min	0	0	0
Mean	72.07	83.80	83.56
SD	(38.61)	(22.87)	(22.70)
Max	100	100	100
N	1,780	1,780	1,820

Table OA.5: Summary Statistics of Deviation from Optimal Weight

Deviation from Optimal Weight	***UA_in_AI***	***BA_in_AI***	***BA_in_Self***
Min	0	0	0
Mean	40.73	18.05	18.14
SD	(29.15)	(20.78)	(20.94)
Max	100	100	100
N	1,780	1,780	1,820

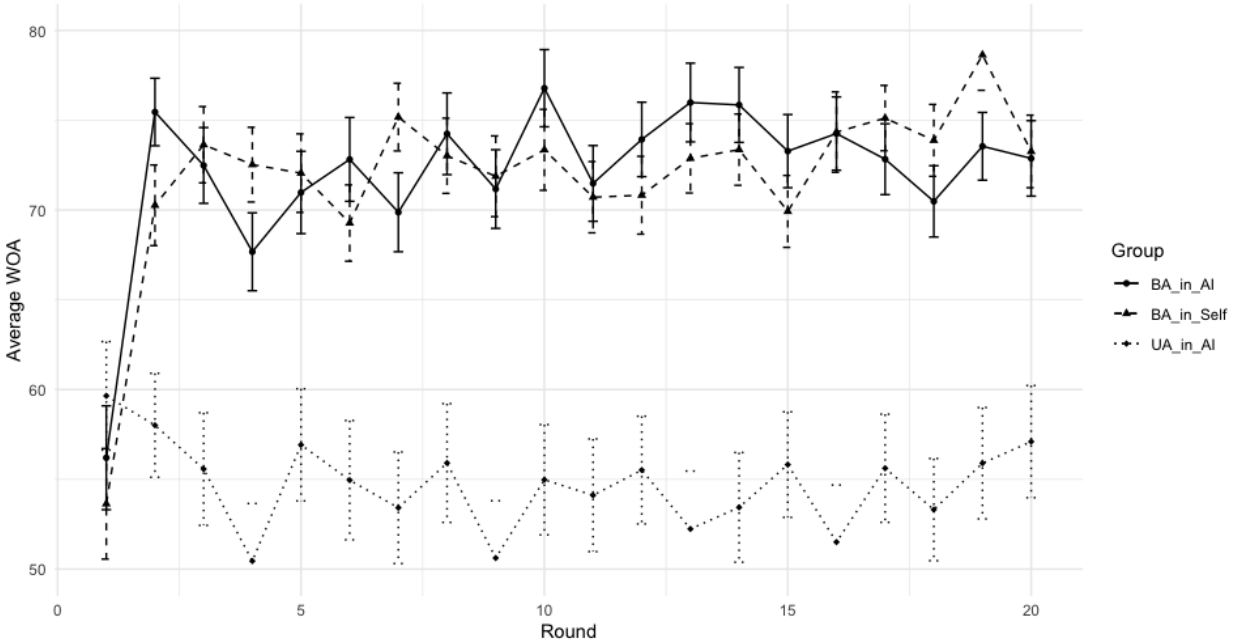
Table **??** presents how people adjust the weight they placed on AI predictions under the *BA_in_Self* condition.

Table OA.6: How Participants Adjust *WOA* Based on Performance Feedback Under *BA_in_Self*

	WOA	
	BA_in_Self	
	(1)	(2)
lag_AE_revised	-0.02^{*}	0.46^{***}
	(0.01)	(0.03)
lag_AE_AI		-0.60^{***}
		(0.03)
Fixed-effects Individual ID	Yes	Yes
N	1729	1729
Adjusted R^2	0.10	0.30

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Standard errors are in parentheses and clustered at an individual level, estimated with individual participant fixed effects.

Figure OA.27: Average *WOA* Across Round Under *BA_in_Self* Condition



Note: Error bars indicate 95% confidence intervals.

Cognitive Effort on Integrating AI Prediction Table OA.7 presents the summary statistics for the time spent on the participants' original prediction, deciding the weights placed on AI or elf predictions, and feedback for each round under each condition.

Table OA.7: Descriptive Statistics of Main Variables

Variable	**Description**	**Treatment**	**Min**	**Mean (SD)**	**Max**	**N**
PT_in_original	Processing time (in seconds) that participants spent on original prediction	*No_AI*	2	14.41 (29.75)	718	1820
		UA_in_AI	3	12.48 (16.21)	246	1780
		BA_in_AI	2	13.68 (22.36)	438	1780
		BA_in_Self	2	11.65 (15.49)	343	1820
PT_in_weight	Processing time (in seconds) the participants spent on weight assignment to AI or original predictions	*No_AI*	1	5.47 (10.07)	308	1820
		UA_in_AI	1	7.82 (8.63)	168	1780
		BA_in_AI	1	9.77 (9.85)	210	1780
		BA_in_Self	1	8.62 (8.25)	117	1820
PT_in_feed	Processing time (in seconds) the participants spent on performance feedback	*No_AI*	1	4.05 (7.67)	212	1820
		UA_in_AI	1	3.54 (5.55)	163	1780
		BA_in_AI	0	3.35 (5.55)	110	1780
		BA_in_Self	0	3.49 (21.81)	907	1820