

The Effect of Interpretable Artificial Intelligence on Repeated Managerial Decision-Making under Uncertainty

Onur Altintas
Boston University
onura@bu.edu

Abraham Seidmann
Boston University
avis@bu.edu

Bin Gu
Boston University
bgu@bu.edu

Nina Mažar
Boston University
nmazar@bu.edu

Abstract

Business decisions involving investments, healthcare, and supply chains are often made in uncertain environments. At the same time, despite being optimal initially, such choices may seem incorrect in hindsight, which may explain why decision-makers hesitate to use AI algorithms under high uncertainty. While some studies suggest that making AI and ML applications more understandable can boost their adoption and trust, this hasn't been examined in uncertain conditions where decision-makers must make repetitive business decisions. Our study addresses this issue empirically by analyzing how different interpretability approaches affect AI adoption and trust under varying levels of uncertainty. Surprisingly, we find that providing interpretability does not necessarily increase AI adoption. In some cases, it may even reduce AI adoption. Interestingly, even though AI adoption was higher, trust in the AI recommendations was significantly lower in high uncertainty compared to low uncertainty across all interpretability types. The evidence is clear that showing the cumulative monetary performance of AI to the users as a benchmark, side by side with their own monetary performance, enhances trust in the AI recommendations.

Keywords: Human-AI Interaction, Interpretability, Uncertainty, Decision-Making

1. Introduction

With advances in machine learning (ML) and artificial intelligence (AI), companies are accelerating plans to implement these systems to improve decision-making, increase revenue, and gain a competitive advantage. Although companies have been integrating AI tools into their processes, decision-makers are often reluctant to accept algorithmic recommendations. One possible reason for resistance is not understanding how the recommendations are generated. AI models are difficult

to fully comprehend and may cause crises (Mayer et al., 2020).

One approach suggested in the literature is to use more inherently interpretable models or to provide interpretability to black-box models post-hoc (Rudin, 2019). Some empirical studies show that interpretability increases reliance on a model and decision-making performance in a predictable environment where the performance measure of the tools is their accuracy (Lai & Tan, 2019; Poursabzi-Sangdeh et al., 2021). In reality, many business decisions are made under high uncertainty, posing challenges for decision-makers. Firstly, they often lack information about the uncertainty level. Secondly, uncertainty complicates decision quality evaluation. An optimal choice may seem wrong due to noisy outcomes. The same action might yield different results each time it's taken. An instance is the classic newsvendor problem, where inventory decisions are made under uncertainty. Despite having a clear optimal choice, decision-makers view it as incorrect and often rely on past demand. Thus, earlier studies haven't addressed the typical challenges uncertainty brings to managerial decision-making, though they demonstrate how interpretability influences advice-taking. Hence, earlier studies haven't captured the typical challenges uncertainty poses for managerial decision-making despite showing that interpretability positively affects advice-taking.

We investigate if interpretability boosts AI adoption and fosters trust in a superior-performing AI model in an uncertain, repetitive business setting. We study the impact of various interpretability types on AI adoption and trust at different uncertainty levels using a lab study. We discover interpretability doesn't always increase adoption or trust; in fact, certain types can diminish them. Specifically, explaining how the AI model decides on the suggested order quantity seems to make users believe they can calculate a more profitable order quantity themselves, especially in less variable demand environments, due to what we call *computational overconfidence*. Surprisingly, higher uncertainty leads

to greater AI adoption but less trust in all interpretability groups, as users' awareness of their limitations seems to increase reliance on the AI tool. Finally, providing cumulative performance data of the AI tool over time boosts trust in both high and low-uncertainty scenarios.

2. Related Studies

It has been shown that increasing the transparency of a model increases the reliance on the model in multiple settings, such as deception detection and clinical diagnosis (Cadario et al., 2021; Lai & Tan, 2019). It is also argued that if users do not trust a model or a prediction, they will not use it (Carvalho et al., 2019). Even the most inherently interpretable models might be hard to understand by their users with possibly little to no technical knowledge. Therefore, there is a certain need to provide a comprehensible explanation to a novice user even if the model is inherently interpretable, such as linear additive models or tree or rule-based models. Glass et al. (2008) show that the availability of explanation capabilities in complex decision support systems could improve user trust. Goodwin et al. (2013) find that in a forecasting task, participants might not adjust their predictions towards advice even though they state higher trust in the model with explanation than in the model with no explanation. Suresh et al. (2020) find that participants who get more information about the ML system follow the system's recommendations more than participants who are only given the recommendations without explanation. These findings lead to our first research question:

RQ1: *How does interpretability affect the adoption and trust in AI models in uncertain environments?*

The amount of information provided might impact user understanding, usage, and trust in a model. In some instances, users might demand more information about the model based on their previous knowledge. There is a tradeoff between providing enough information to promote understanding and to avoid information overload. Gönül et al. (2006) show that long explanations and explanations with higher information value are more effective. Kulesza et al. (2013) find that greater completeness of explanations promotes user trust. Bussone et al. (2015) argue that detailed explanations increase reliance and enhance trust in the system. Accordingly, we investigate the question:

RQ2: *Do more comprehensive explanations affect the adoption and trust in AI models in uncertain environments?*

Testing different types of explanation is especially important under high uncertainty since the effect of such heuristics and biases amplify when making

decisions under uncertainty (Tversky & Kahneman, 1974). Interpretability requires explanation methods to convey information about the system to the end user. The literature mainly classifies explanations based on the nature of the queries, such as "Why" and "How". "Why" and "How" explanations were first introduced by MYCIN, an artificial intelligence program to treat blood infections that can explain the reasoning that led to its diagnosis and recommendation (Buchanan and Shortliffe, 1984). Dhaliwal and Benbasat (1996) shows that the use of the "Why" and "How" explanations resulted in higher perceptions of usefulness than the use of other explanation types. Lim et al. (2009) find that "Why" and "Why Not" explanations improve understanding and task performance and increase trust in the system. W. Wang and Benbasat (2007) show how explanations increased user trust for an e-commerce recommender system. Haynes et al. (2009) constructed an explanation framework for intelligent agent systems by synthesizing previous studies. They propose four categories of explanations: Design (Why), Mechanistic (How), Ontological (What), and Operational. Based on the previous studies, we incorporate the two most common explanation queries in our study: "Why" and "How".

RQ3: *What are the effects of design ("why") explanations and mechanistic ("how") explanations on the adoption and trust in AI models in uncertain environments?*

Uncertainty is an important element of decision-making and advice-taking. Highly uncertain environments complicate the decision-making process and make it harder to evaluate any advice. To the best of our knowledge, there are only a few studies that capture the effect of uncertainty on advice-taking, yet these studies do not capture the effect of interpretability. Dietvorst and Bharti (2020) investigates the decision-making behavior of users under uncertainty. They find that users resist advice from an algorithm more when the environment is more uncertain. D. Wang et al. (2021) find that communicating the level of uncertainty to users does not affect trust, confidence, or decision-making quality, but users spend more time making decisions when they receive information about uncertainty. In their work, they use uncertainty as the level of input certainty captured by the model rather than irreducible environmental uncertainty.

In a more general context, it has been shown that uncertainty leads people to seek more information (Brashers & Hogan, 2013; Kenny et al., 2021). If explaining a model promotes AI usage and trust in AI, then the effect of the explanation should be

greater in a more uncertain environment. Therefore, we expect the impact of interpretability to increase in such environments (i.e., the interaction effect of interpretability and uncertainty). This leads to the last research question:

RQ4: *How does the change in the level of uncertainty impact the effect of interpretability on AI usage and users' trust in AI?*

3. Experimental Setting and Design

We designed a laboratory experiment to capture the combined effect of different interpretability types of a model and the different levels of environmental uncertainty on AI adoption and trust. In this section, we first describe the user task and the data collected.

3.1. Experiment Task

We used the newsvendor problem as our testing ground, where a manager repeatedly orders perishable goods for later sale, making decisions before knowing the demand. Though extensively studied, our focus isn't on the decision-making aspect but rather on assessing AI adoption and trust using this familiar scenario.

In our experiment, participants are asked to act as bread store managers, deciding how many loaves to order daily, aiming to maximize cumulative profit. They were given cost parameters and instructions on calculating daily profit based on actual sales and informed that demand varies randomly. Unlike the traditional newsvendor problem, our setting reflects a more realistic business setting where the underlying demand distribution for a product is not available to the managers. Therefore, in our task, the underlying demand distribution was not shared with the participants so that they could self-evaluate the level of uncertainty in demand.

The treatment groups received order quantity recommendations from the adaptive inventory management (AIM) algorithm Huh and Rusmevichientong (2009). We selected this algorithm as it quickly converges to the optimal order quantity. In the experiment, the algorithm's recommendations were about 3% away from the optimal quantities in the final five rounds. The algorithm's dynamic tree structure, which adjusts recommendations using prior sales data, makes it inherently interpretable and easy to explain to novice users. Additionally, it utilizes the online gradient descent approach, frequently used in online machine learning (Shalev-Shwartz & Ben-David, 2014, p.300).

3.2. Experiment Treatments

The experiment focuses on the effect of model interpretability and environmental uncertainty. Therefore, we created a 2×6 between-subject factorial experiment design. The experiment groups are as follows:

1. **Uncertainty (2 levels):** We created two levels of environmental uncertainty, high and low, by changing the demand distribution parameters. The daily demand sequence is generated from $U(50, 950)$ for the high-uncertainty and $U(450, 550)$ for the low-uncertainty group.
2. **Information and Support (6 levels):** The experiment includes one control and five treatment groups, which differ in terms of the additional information and decision support they receive about the AI model (black-box, design, mechanistic, comprehensive, and performance feedback). As there are various ways to provide additional information about the models to make them more interpretable, we followed the taxonomies and the frameworks in the literature to construct two textual-based, model-specific explanations (Dodge et al., 2018; Haynes et al., 2009). Using two model-specific explanations, we created three different interpretability types (design, mechanistic, and comprehensive).

In the experiment, six groups were provided with varying AI support and information. The control group had no AI support or additional information. The black-box group had AI support and general AI information without specifics on the model. The design group was given AI support and details on the AI's design logic, which focuses on the reasoning behind the recommendations (i.e., why the AI model is doing what it is doing along with its capabilities). The mechanistic group received AI support and information on the AI's inner workings, where participants can observe how the AI model generates its recommendation step by step in a decision tree format. The comprehensive group combined the design and mechanistic groups, providing AI support with both design logic and inner workings. Lastly, the performance feedback group could observe the continuous cumulative monetary performance of the AI tool side-by-side with their own cumulative monetary performance.

3.3. Experiment Flow

Upon enrolling in the study, participants are given an overview of its length, rewards, and objectives, and

they must provide informed consent to participate. They are then familiarized with the experimental setup, which includes business metrics, demand data, and how profits are computed. Participants also share their initial level of trust in AI within business environments, expressed on a 7-point Likert scale, and respond to a series of attention-check questions.

Subsequently, they are randomly allocated to different groups, each facing varying levels of uncertainty and receiving distinct treatments. Participants initially engage in three practice rounds to get accustomed to the experiment. These are followed by five main rounds where they are introduced to an AI tool. Depending on the group they are assigned to, participants receive varying degrees of information about the AI tool. We provided a set of comprehension questions to ensure the understanding of explanations by participants.

Over the course of 30 decision-making rounds, those in the treatment groups receive order suggestions from the AI tool after making their own initial order estimates. They can choose to either stick with their own decision or adopt the AI's recommendation. This decision-making model is inspired by the judge-advisor system. Additionally, certain groups receive real-time feedback regarding the cumulative profits achieved by the AI's recommendations.

In the final 10 rounds, all participants make order decisions without any input from the AI tool. The experiment concludes with participants completing a survey that collects demographic information and gauges their trust in AI on a 7-point Likert scale, both in the context of the experiment and in broader business scenarios. Participants are then debriefed on their compensation for taking part in the study.

Additionally, participants received two attention-testing questions in rounds 11, 26, and 41. They receive an additional bonus payment for each correct answer. Figure 1 shows the flow of the experiment.

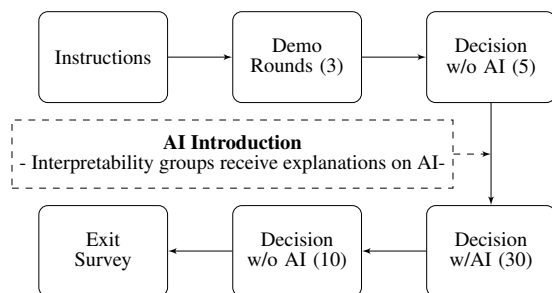


Figure 1. The flow of the experiment. Numbers in parentheses are the number of rounds

4. Experimental Study

We recruited 642 participants from Amazon Mechanical Turk (Mturk) in May 2022¹. We used the Cloudresearch platform to filter low-quality participants to ensure data quality. Our experiment was available only to college graduates with age between 20 and 47 years old who had completed at least 1,000 HITs with a greater than 95% approval rate. Participants were compensated US\$ 2.75 for completing the experiment in around 30 minutes. Additionally, participants could earn a bonus proportional to their cumulative profit performance throughout the experiment and correct answers to the attention questions. Out of 642 participants who completed the experiment, 33 were excluded from the analyses due to their low performance (less than half) in the attention questions, extremely low or high order quantities over multiple rounds, or incomprehensible exit survey answers. Ultimately, 609 participants were included in the analyses (342 female, 260 male, and 7 identified as other, average age = 34.6). The participants were paid US\$ 4.90 on average for the experiment, which took them about 27.2 minutes to complete.

The premise of our study is that decision-makers do not follow better-performing AI models when they should. If the participants can perform better than the algorithm, then the resistance to using the algorithm more often will be reasonable. However, this is not the case in the study. AI's cumulative profit between round 6 and round 35 was significantly higher than the control and any information group in both uncertainty levels (all pairwise t-test p-values <0.0001). Since the experiment's setting has inherent uncertainty, participants may perform better than the AI model. Still, we expect a low number of participants to outperform the model. In line with our expectations, only 52 out of 609 participants (8.5%) earned more profit than the model (23 participants in the low-uncertainty group and 29 participants in the high-uncertainty group). The participants could have earned about 10% more profit on average if they had followed the AI recommendations repeatedly (3% for the low and 27% for the high-uncertainty group). The statistics of the main variables for the study are reported in Table 1 for each uncertainty level.

In the following two sections, we investigate how AI adoption and trust in AI are affected by the various interpretability types and the level of environmental uncertainty.

¹142 performance feedback group participants were recruited in September 2022.

Table 1. Summary statistics of the treatment groups for each uncertainty level

| | Mean | SD | N |
|-------------------------|-------|-------|-----|
| <i>Low Uncertainty</i> | | | |
| Avg AI adoption | 37.5% | 0.270 | 253 |
| Ex-ante trust | 4.41 | 1.399 | 253 |
| Ex-post trust | 4.82 | 1.521 | 253 |
| <i>High Uncertainty</i> | | | |
| Avg AI adoption | 48.3% | 0.273 | 263 |
| Ex-ante trust | 4.47 | 1.313 | 263 |
| Ex-post trust | 3.82 | 1.746 | 263 |

Notes: The statistics in the table are aggregate values for the treatment groups under respective uncertainty levels. The trust levels are on the 7-Likert scale. The average AI adoption represents the percentage of rounds in which a participant followed AI's recommended order quantity.

4.1. Results

4.1.1. AI Adoption One of the main motivations for providing interpretability is to promote AI adoption. The percentage of the time that participants followed the AI recommendations is reported in Table 2.

Table 2. Summary statistics for average AI adoption per uncertainty and interpretability group

| V: Avg AI Adoption | Mean | SD | N |
|-------------------------|-------|-------|-----|
| <i>Overall</i> | | | |
| Black-box | 43.0% | 0.277 | 516 |
| Design | 45.2% | 0.282 | 95 |
| Mechanistic | 46.5% | 0.269 | 98 |
| Comprehensive | 39.9% | 0.281 | 95 |
| Performance Feedback | 35.7% | 0.264 | 101 |
| <i>Low Uncertainty</i> | | | |
| Black-box | 48.2% | 0.271 | 127 |
| Design | 37.5% | 0.270 | 253 |
| Mechanistic | 40.1% | 0.296 | 46 |
| Comprehensive | 41.7% | 0.276 | 47 |
| Performance Feedback | 29.9% | 0.244 | 48 |
| <i>High Uncertainty</i> | | | |
| Black-box | 33.5% | 0.246 | 50 |
| Design | 41.2% | 0.274 | 62 |
| Mechanistic | 48.3% | 0.273 | 263 |
| Comprehensive | 50.1% | 0.262 | 49 |
| Performance Feedback | 50.8% | 0.257 | 51 |
| Black-box | 46.1% | 0.295 | 47 |
| Design | 37.8% | 0.259 | 51 |
| Mechanistic | 54.8% | 0.253 | 65 |
| Comprehensive | | | |
| Performance Feedback | | | |

Note: Statistics of average AI adoption shared above are at the participant level

The difference in average AI adoption between the interpretability groups is significant in our ANOVA tests ($F_{4,511} = 4.31, p = .0019$). This also holds when we control for the level of uncertainty ($F_{4,510} = 3.7,$

$p = .0018$). Pairwise two-sided t-tests show that providing interpretability might negatively affect AI usage for some groups. Both the mechanistic and the comprehensive groups followed AI recommendations significantly less than the black-box group ($p = .074$ and $p = .015$, respectively). The mean AI adoption rate in the black-box group is similar to that in both the design group ($p = .76$) and the performance feedback group ($p = .44$).

When we compare the effect of different explanations on AI adoption, we see that the design group had higher usage than the mechanistic and comprehensive groups ($p = .032$ and $p = .005$, respectively). Similarly, the performance feedback group followed the AI recommendations significantly more than the mechanistic and the comprehensive groups ($p = .0067$ and $p = .0005$, respectively).

Table 3. The effect of interpretability, uncertainty, and ex-ante AI trust on AI usage

| LPM | DV: AI Usage | |
|---------------------------------------|--------------|----------|
| | (1) | (2) |
| Interpretability | | |
| Black-box (Reference) | 0.250*** | 0.252*** |
| | (0.050) | (0.058) |
| Design | 0.017 | 0.022 |
| | (0.038) | (0.057) |
| Mechanistic | -0.067+ | -0.097+ |
| | (0.039) | (0.055) |
| Comprehensive | -0.087* | -0.055 |
| | (0.038) | (0.055) |
| Performance Feedback | 0.027 | 0.004 |
| | (0.036) | (0.053) |
| Uncertainty | | |
| High | 0.105*** | 0.098+ |
| | (0.023) | (0.055) |
| Interpretability × Uncertainty | | |
| Design × High | | -0.009 |
| | | (0.076) |
| Mechanistic × High | | 0.059 |
| | | (0.078) |
| Comprehensive × High | | -0.062 |
| | | (0.076) |
| Performance Feedback × High | | 0.044 |
| | | (0.071) |
| Ex-ante AI Trust | 0.033*** | 0.033*** |
| | (0.009) | (0.009) |
| Adj. R ² | 0.029 | 0.031 |
| F-stat | 10.205 | 6.546 |
| Observations | 15480 | 15480 |

*Note: The dependent variable (AI usage) is a binary variable whether a participant chose the AI recommendation as the final order quantity. Ex-ante AI trust is on a 7-Likert scale. Only rounds between 6 and 35, where the AI support was available, are included in the regression. The black-box group with low uncertainty is the reference group. (1) and (2) are linear probability models. Participant-level clustered standard errors are in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

Additionally, we conduct regression analyses, and the output is reported in Table 3. The dependent variable of the models is AI usage, which indicates

whether a participant followed the AI recommendation in a given period. The independent variables of the models are interpretability types, uncertainty level of the environment, and ex-ante AI trust. The baseline group is the black-box group under low uncertainty in all regression models. We analyze the effect of interpretability and the uncertainty level using a linear probability model (Table 3, Model 1). We found that the mechanistic and the comprehensive groups followed the AI recommendation significantly less often than the black-box group (by about, on average, 6.8% and 8.8%, respectively). Both the design group and the performance feedback group used AI recommendations more than the black-box group, but the difference is insignificant.

The literature suggests that users rely more on their own decisions when uncertainty increases (Dietvorst and Bharti, 2020). Our regression analyses show the contrary: AI adoption increases as the level of uncertainty increases. The participants in the high-uncertainty environment followed AI's recommendations on average 10% more often than those in the low-uncertainty environment. In Table 3, Model 2, we added the interaction effect of the interpretability types and the level of uncertainty. Interestingly, the higher level of uncertainty diminished the effect of interpretability (opposite interaction and main effect estimates) for the design and the mechanistic groups, whereas it amplified the effect for the comprehensive and the performance feedback groups. These effects are robust in the logistic regression as well.

AI Adoption over Time In this section, we look at how interpretability affects the AI usage trend over time with different uncertainty levels. First, we analyze the AI adoption behavior of participants over time using multiple linear probability models where the dependent variable is AI adoption, which indicates whether a participant followed the AI recommendation in a given period. The independent variables of the models are the interpretability type, the uncertainty level of the environment, ex-ante AI trust, and round. The baseline group is the black-box group with low uncertainty in all regression models.

We find that the participants decreased their AI usage over time in general (Table 4, Model 1). However, we see that this negative impact was suppressed for the interpretability groups (Model 2). The AI usage trend for the treatment groups was positive (Models 3-4). Every ten rounds, the AI usage probability of those participants increases by 4% on average.

The negative AI usage trend is driven by the participants in the high-uncertainty environment. They

followed the AI recommendations more often (on average 33%) than those in the low-uncertainty environment. However, their AI usage decreased on average by around 1.1% every round (Table 4, Models 3-4).

Table 4. The effect of interpretability and uncertainty on the AI usage trend

| LPM | DV: AI Usage _t | | | |
|---------------------------------------|--------------------------------|--------------------------------|----------------------|----------------------|
| | (1) | (2) | (3) | (4) |
| Black-box (Reference) | 0.273*** (0.051) | 0.158** (0.050) | 0.225*** (0.054) | 0.227*** (0.060) |
| Round | -0.001 ⁺ (0.001) | 0.004*** (0.001) | 0.001 (0.001) | 0.001 (0.001) |
| Interpretability | | | | |
| Design | 0.017 (0.038) | 0.017 (0.038) | -0.069 (0.052) | -0.064 (0.066) |
| Mechanistic | -0.067 ⁺ (0.039) | -0.067 ⁺ (0.057) | -0.124* (0.057) | -0.153* (0.060) |
| Comprehensive | -0.087* (0.038) | -0.087* (0.053) | -0.176*** (0.053) | -0.145* (0.059) |
| Performance Feedback | 0.027 (0.036) | 0.027 (0.036) | -0.066 (0.046) | -0.088 (0.060) |
| Uncertainty | | | | |
| High | 0.105*** (0.023) | 0.332*** (0.031) | 0.332*** (0.031) | 0.324*** (0.058) |
| Uncertainty × Round | | | | |
| High × Round | | -0.011*** (0.001) | -0.011*** (0.001) | -0.011*** (0.001) |
| Interpretability × Round | | | | |
| Design × Round | | | 0.004* (0.002) | 0.004* (0.002) |
| Mechanistic × Round | | | 0.003 (0.002) | 0.003 (0.002) |
| Comprehensive × Round | | | 0.004* (0.002) | 0.004* (0.002) |
| Performance Feedback × Round | | | 0.005* (0.002) | 0.005* (0.002) |
| Interpretability × Uncertainty | | | | |
| Design × High | | | | -0.009 (0.076) |
| Mechanistic × High | | | | 0.059 (0.079) |
| Comprehensive × High | | | | -0.062 (0.076) |
| Performance Feedback × High | | | | 0.044 (0.071) |
| Ex-ante AI Trust | 0.033*** (0.009) | 0.033*** (0.009) | 0.033*** (0.009) | 0.033*** (0.009) |
| Adj. R ² | 0.030 | 0.039 | 0.040 | 0.041 |
| F-stat | 8.92 | 22.43 | 15.36 | 11.87 |
| Observations | 15480 | 15480 | 15480 | 15480 |

*Note: The dependent variable is a binary variable that indicates whether a participant chose AI's recommendation as a final order quantity in a given round. Ex-ante AI trust is on a 7-Likert scale. Rounds 6 to 35 are included in the regression where AI support was available. The black-box group in the low-uncertainty environment is the reference group. Participant-level clustered standard errors are in parentheses. ⁺ p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001*

4.1.2. Trust in AI At the end of the experiment, participants reported their level of trust in AI on a 7-Likert scale, with 1 indicating the lowest and 7 the highest level of trust. In the following analyses, we use the end-of-experiment AI trust (ex-post AI trust) as our main metric. The descriptive statistics for the ex-post AI trust are reported in Table 5.

The effect of uncertainty on AI trust is highly significant in our experiment (ANOVA $F(1, 514) = 48.06, p < 0.0001$). Participants in

Table 5. Summary statistics for self-reported ex-post trust in AI by interpretability groups at different uncertainty levels

| V: Ex-post Trust | Mean | SD | N |
|-------------------------|------|------|-----|
| <i>Overall</i> | 4.31 | 1.71 | 516 |
| Black-box | 4.09 | 1.83 | 95 |
| Design | 4.19 | 1.76 | 98 |
| Mechanistic | 3.93 | 1.78 | 95 |
| Comprehensive | 4.33 | 1.67 | 101 |
| Performance Feedback | 4.84 | 1.45 | 127 |
| <i>Low Uncertainty</i> | 4.82 | 1.52 | 253 |
| Black-box | 4.74 | 1.58 | 46 |
| Design | 4.81 | 1.61 | 47 |
| Mechanistic | 4.33 | 1.72 | 48 |
| Comprehensive | 4.88 | 1.49 | 50 |
| Performance Feedback | 5.23 | 1.15 | 62 |
| <i>High Uncertainty</i> | 3.82 | 1.75 | 263 |
| Black-box | 3.49 | 1.85 | 49 |
| Design | 3.63 | 1.71 | 51 |
| Mechanistic | 3.51 | 1.76 | 47 |
| Comprehensive | 3.78 | 1.68 | 51 |
| Performance Feedback | 4.48 | 1.61 | 65 |

Notes: The ex-post AI trust is on a 7-Likert scale per participant collected at the end of the experiment

the high-uncertainty environment trust AI less than those in the low-uncertainty environment by an average of 1 Likert point (Model 1 in Table 6).

On the other hand, the treatment groups reported different levels of ex-post AI trust. The performance feedback group reported significantly higher trust levels than any other treatment group at both uncertainty levels. The ex-post AI trust for the comprehensive group is significantly higher than the black-box group only when we control AI usage (Model 2 in Table 6). We also found that the negative effect of increasing uncertainty on ex-post AI trust is smaller for the interpretability groups compared to the black-box group (Model 3 in Table 6).

We can expect that self-reported ex-post AI trust can be affected by multiple factors. For example, a user with higher AI adoption might have a different trust level than a user with lower AI adoption. Additionally, a user with low ex-ante trust in AI might have lower ex-post trust in AI. When we control for ex-ante AI trust, we see that the reported ex-ante and ex-post AI trust are positively correlated (Table 6). Thus, providing interpretability can affect ex-post AI trust through AI usage, which suggests a possible mediation effect of AI usage on ex-post AI trust.

Table 6. The regression model for the effect of interpretability, uncertainty, and ex-ante AI trust on ex-post AI trust

| OLS | DV: Ex-post Trust | | |
|---------------------------------------|----------------------|----------------------|----------------------|
| | (1) | (2) | (3) |
| Black-box (Reference) | 3.012*** (0.324) | 2.267*** (0.273) | 2.375*** (0.296) |
| Interpretability | | | |
| Design | 0.164 (0.227) | 0.112 (0.191) | 0.068 (0.259) |
| Mechanistic | -0.149 (0.244) | 0.052 (0.202) | -0.059 (0.279) |
| Comprehensive | 0.303 (0.229) | 0.561** (0.187) | 0.421+ (0.254) |
| Performance Feedback | 0.718*** (0.203) | 0.637*** (0.179) | 0.410+ (0.226) |
| Uncertainty | | | |
| High | -1.026*** (0.135) | -1.339*** (0.117) | -1.558*** (0.275) |
| Interpretability × Uncertainty | | | |
| Design × High | | | 0.088 (0.381) |
| Mechanistic × High | | | 0.214 (0.407) |
| Comprehensive × High | | | 0.273 (0.375) |
| Performance Feedback × High | | | 0.443 (0.358) |
| Ex-ante AI Trust | 0.356*** (0.057) | 0.259*** (0.049) | 0.260*** (0.049) |
| Avg AI Usage | | 2.979*** (0.211) | 2.972*** (0.211) |
| Adj. R ² | 0.192 | 0.403 | 0.401 |
| F-stat | 25.168 | 61.074 | 41.199 |
| Observations | 516 | 516 | 516 |

Note: The table shows the output of OLS regressions on ex-post AI trust (7-Likert scale) per participant at the participant level. Ex-ante AI trust is on 7-Likert scale. The baseline is the black-box group with low uncertainty. Participant-level-clustered standard errors are in parentheses.

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5. Conclusion

AI-based sophisticated decision support systems have become more common, but decision-makers tend to resist adopting algorithmic advice consistently for many different tasks and settings. It is argued that one of the main reasons for the resistance is the decision-makers' inability to understand how the model generates its recommendations. A possible solution offered in the literature is to provide model interpretability. While several recent studies claim that interpretability potentially increases the adoption of and trust in sophisticated models, these claims have not been tested in a setting with repetitive managerial decisions under uncertainty.

Most business decisions are made in an uncertain environment. Uncertainty creates multiple challenges.

Due to the lack of information and the noisy outcome, cause-and-effect relationships become obscure. In our study, we investigate whether interpretability can help decision-makers overcome such challenges, increasing AI adoption and trust when they make repetitive business decisions under different levels of uncertainty.

5.1. General Discussion

The results of our experiment show that providing interpretability does not promote AI adoption in an uncertain environment. Interestingly, we find that explaining the inner mechanism of a model decreases AI adoption. In our opinion, the resistance to adoption is due to *computational overconfidence*. Based on our participants' qualitative feedback, we found that explaining the inner mechanism helps participants understand how the model works and the input-output relationship, which allows them to simulate the recommendation themselves. The ability to simulate the algorithm's recommendations leads participants to mistakenly believe that they can perform better than the model. However, such *computational overconfidence* might lead to insufficient utilization of better-performing decision-support models. Similar resistance to an interpretable algorithm in the context of repetitive decision-making under uncertainty has been recently reported by DeStefano et al. (2022). Their novel field experiment was conducted at a retail company while it was switching its stock reordering decision support system from an interpretable (weighted moving average) model to an uninterpretable (recurrent neural network) model. They have shown that the decision-makers had higher usage of the uninterpretable model as compared with the former interpretable model. Follow-up qualitative interviews in this study revealed that managers use interpretability information to confirm their prior opinions, which lowers the acceptance of the AI recommendations. These results seem to further reinforce the evidence for computational overconfidence.

The literature suggests that users rely more on their own decisions when uncertainty increases (Dietvorst & Bharti, 2020). Our regression analyses show the contrary: AI adoption increases as the level of uncertainty increases. The participants in the high-uncertainty environment followed AI's recommendations more often than those in the low-uncertainty environment. Yet, surprisingly, those in the high-uncertainty environment reported lower trust in AI. This suggests that AI adoption and self-reported trust in AI can produce different results (Papenmeier et al., 2022; Rechkemmer & Yin, 2022). While making

repetitive decisions in uncertain environments, it is harder for participants to make decisions confidently due to the obscure nature of the cause-and-effect relationship. This not only decreases the participants' trust in their decision but also causes additional cognitive load (Rad & Pham, 2017). This might suggest that participants have higher confidence in the model's recommendation than their own. Thus, they follow its recommendations even though they do not trust it fully (Thompson, 2017).

We also find that an increase in the level of uncertainty decreases AI adoption over time. In uncertain environments, it is harder for decision-makers to evaluate not only their own performance but also the performance of AI tools. We conjecture that participants have (wrongly) used the model's accuracy to evaluate its long-run economic performance. Clearly, its objective is not to predict the next period's demand but rather to maximize the long-run cumulative profit. Thus, our results support that participants evaluate the observed accuracy as the main performance measure of AI, and they decreased their usage over time due to perceived low-performance (Saragih & Morrison, 2022; Yu et al., 2016).

Uncertain environments create unique challenges for AI interpretability and its effectiveness. The results of our study clearly show that providing interpretability is insufficient under high uncertainty. Continuously sharing the performance outcomes of the AI tool, along with the decision-makers' own performance, promotes AI trust without any resistance to adoption. This indicates that providing both interpretability and human versus AI real-time performance can significantly improve both AI usage and trust in AI.

5.2. Limitations

Our study examines the relationship between interpretability and irreducible environmental uncertainty for a repetitive business task. We aimed for a common task, the classic newsvendor problem, to ensure the generalizability of our results. Although we think our results will be carried over to most business settings, participants might exhibit different behaviors in other specific tasks.

The use of tools based on sophisticated AI models, especially by novice users, will undeniably expand as these tools become increasingly ubiquitous. We designed our experiment settings specifically for novice users recruited on Amazon Mechanical Turk (Mturk). Although there is controversy about the suitability of Mturk participants in behavioral experiments, multiple studies show that Mturk participants provide

high-quality data that better represents the general population than typical student subjects commonly used in behavioral studies (Chandler et al., 2019; Lee et al., 2018). Additionally, we filtered out our participants using the Cloudresearch platform, whose participant pool has proven to represent the general population better and generate higher quality data than the typical Mturk participant pool (Chandler et al., 2019; Hauser et al., 2022).

Lastly, we chose to use a decision tree-based model, which is considered to be one of the most common inherently interpretable models. To provide high-quality explanations, we followed the explanation theory literature. Even though we think that our results are robust to model and explanation changes, different models and explanations might impose additional effects that might impact our findings.

References

- Brashers, D. E., & Hogan, T. P. (2013). The appraisal and management of uncertainty: Implications for information-retrieval systems. *Information Processing & Management*, 49(6), 1241–1249.
- Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems : The MYCIN experiments of the Stanford Heuristic Programming Project*. Reading, Mass. : Addison-Wesley.
- Bussone, A., Stumpf, S., & O’Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics*, 160–169.
- Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, 5(12), 1636–1642.
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics [Number: 8 Publisher: Multidisciplinary Digital Publishing Institute]. *Electronics*, 8(8), 832.
- Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51(5), 2022–2038.
- DeStefano, T., Kellogg, K., Menietti, M., & Vendraminelli, L. (2022). Why Providing Humans with Interpretable Algorithms May, Counterintuitively, Lead to Lower Decision-making Performance. *SSRN Electronic Journal*.
- Dhaliwal, J. S., & Benbasat, I. (1996). The Use and Effects of Knowledge-based System Explanations: Theoretical Foundations and a Framework for Empirical Evaluation [Publisher: INFORMS]. *Information Systems Research*, 7(3), 342–362.
- Dietvorst, B. J., & Bharti, S. (2020). People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error. *Psychological Science*, 31(10), 1302–1314.
- Dodge, J., Penney, S., Anderson, A., & Burnett, M. (2018). What Should Be in an XAI Explanation? What IFT Reveals, 4.
- Glass, A., McGuinness, D. L., & Wolverton, M. (2008). Toward establishing trust in adaptive agents. *Proceedings of the 13th international conference on Intelligent user interfaces*, 227–236.
- Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems*, 42(3), 1481–1493.
- Goodwin, P., Sinan Gönül, M., & Önkal, D. (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting*, 29(2), 354–366.
- Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2022). Evaluating CloudResearch’s Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods*.
- Haynes, S. R., Cohen, M. A., & Ritter, F. E. (2009). Designs for explaining intelligent agents. *International Journal of Human-Computer Studies*, 67(1), 90–110.
- Huh, W. T., & Rusmevichientong, P. (2009). A Nonparametric Asymptotic Analysis of Inventory Planning with Censored Demand. *Mathematics of Operations Research*, 34(1), 103–123.
- Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 294, 103459.
- Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users’ mental models [ISSN: 1943-6106]. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10.

- Lai, V., & Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection [arXiv: 1811.07901]. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.
- Lee, Y. S., Seo, Y. W., & Siemsen, E. (2018). Running Behavioral Operations Experiments Using Amazon’s Mechanical Turk. *Production and Operations Management*, 27(5), 973–989.
- Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2119–2128.
- Mayer, A.-S., Strich, F., & Fiedler, M. (2020). Unintended Consequences of Introducing AI Systems for Decision Making [Publisher: Association for Information Systems]. *MIS Quarterly Executive*, 19(4), 239–257.
- Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It’s Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction*, 29(4), 1–33.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability [arXiv: 1802.07810]. *arXiv:1802.07810 [cs]*.
- Rad, A. F., & Pham, M. T. (2017). Uncertainty Increases the Reliance on Affect in Decisions. *Journal of Consumer Research*, ucw073.
- Rechkemmer, A., & Yin, M. (2022). When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature machine intelligence*, 1(5), 206–215.
- Saragih, M., & Morrison, B. W. (2022). The Effect of past Algorithmic Performance and Decision Significance on Algorithmic Advice Acceptance. *International Journal of Human-Computer Interaction*, 38(13), 1228–1237.
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms* (1st ed.). Cambridge University Press.
- Suresh, H., Lao, N., & Liccardi, I. (2020). Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. *12th ACM Conference on Web Science*, 315–324.
- Thompson, C. (2017). Trust without Reliance. *Ethical Theory and Moral Practice*, 20(3), 643–655.
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases [Publisher: American Association for the Advancement of Science]. *Science*, 185(4157), 1124–1131.
- Wang, D., Zhang, W., & Lim, B. Y. (2021). Show or Suppress? Managing Input Uncertainty in Machine Learning Model Explanations [arXiv: 2101.09498]. *arXiv:2101.09498 [cs]*.
- Wang, W., & Benbasat, I. (2007). Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. *Journal of Management Information Systems*, 23(4), 217–246.
- Yu, K., Berkovsky, S., Conway, D., Taib, R., Zhou, J., & Chen, F. (2016). Trust and Reliance Based on System Accuracy. *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 223–227.