MODEL BLINDNESS: INVESTIGATING A MODEL-BASED ROUTE-RECOMMENDER SYSTEM'S IMPACT ON DECISION MAKING

A Dissertation
Presented to
The Academic Faculty

by

Sweta Parmar

In Partial Fulfillment
of the Requirements for the degree
DOCTOR OF PHILOSOPHY in Engineering Psychology in the
SCHOOL OF PSYCHOLOGY

Georgia Institute of Technology
December, 2022

MODEL BLINDNESS: INVESTIGATING A MODEL-BASED ROUTE-RECOMMENDER SYSTEM'S IMPACT ON DECISION MAKING

Dissertation committee:

Dr. Rick Thomas, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Karen Feigh
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Jamie Gorman
School of Psychology
Georgia Institute of Technology
Human Systems Engineering
Arizona State University

Dr. Elizabeth Whitaker
Georgia Tech Research Institute (GTRI)

Dr. Sashank Varma
School of Psychology
Georgia Institute of Technology

Date Approved: Dec 2, 2022

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to Dr. Rick Thomas for his invaluable advice, continuous support, and patience. I learned much about decision-making from him and enjoyed learning from all our long coding sessions. I would like to thank Dr. Karen Feigh for serving on both my master's and Ph.D. committees, her detailed feedback from the beginning of my grad school career has always helped me shape my ideas better and helped me better communicate my work. I would like to thank Dr. Jamie Gorman for his invaluable feedback throughout grad school and for being an awesome instructor for all our EP classes. I would also like to thank Dr. Sashank Varma for his technical support and knowledge, for introducing me to XAI literature, and for all the great questions he asked that helped me conceptualize my idea better and come up with this dissertation study. I would also like to thank Dr. Elizabeth Whitaker for giving me initial directions to develop this dissertation study; my meeting with her and Ethan helped me come up with the idea for using a navigation task and a route recommender system for this study. I would also like to thank our collaborator, Dr. David Illingworth, for his constant input in study design and analyses, for being an awesome labmate, and for transferring all his grad school wisdom to me. Thanks to undergraduate research assistants- Ryan, Usayd, Jackie, and Manasi, without them, it would not have been possible to collect participant data and code qualititative data in such a short time frame. Thanks to all the undergraduate research participants of this study.

I would like to offer my special thanks to Dr. Frank Durso for always being an excellent mentor throughout grad school and giving his invaluable advice on research and major life decisions. I would like to thank all members of the Decision Processes

Laboratory and the School of Psychology for their constant support over the last five years. Also, a shout out to all my seniors who have always motivated me with their grad school wisdom and advice– Joel, Ashley, Angela, Brittany, Rachel, Eric, Sadaf, David G., and Terri. Thanks to all the administrative staff of the School of Psychology- Kristie, Shebbie, Sandra, Kaysha, Emily, Tikica, and Arian for always making tasks easier.

Thanks to my parents for their unwavering support for my academic endeavors when I decided to move to another country to pursue a Ph.D. Thanks, Dewansh, for always supporting and encouraging me and reminding me that I am doing my best. Thanks to my brother, sister, and friends- Vaibhav, Gaurav, Sanchayni, Abhinav, and everyone else who were always there to hear me rant about grad school and human-subject research and continuously encouraging me to continue giving my best.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF SYMBOLS AND ABBREVIATIONS

DSS	Decision-Support Systems
MDSS	Model-based Decision-Support Systems
TMPLAR	Tool for Multi-objective Planning and Asset Routing
XAI	Explainable Artificial Intelligence
BS	Brier Score

SUMMARY

Model-Based Decision Support Systems (MDSS) are prominent in many professional domains of high consequence, such as aeronautics, emergency management, military command and control, healthcare, nuclear operations, intelligence analysis, and maritime operations. An MDSS generally uses a simplified model of the task and the operator to impose structure to the decision-making situation and provide information cues to the operator that is useful for the decision-making task. Models are simplifications, can be misspecified, and have errors. Adoption and use of these errorful models can lead to the impoverished decision-making of users. I term this impoverished state of the decision-maker ***model blindness***. A series of two experiments were conducted to investigate the detrimental consequences of model blindness on human decision-making and performance and how those consequences can be mitigated via an explainable AI (XAI) intervention. This dissertation also reports simulation results that motivated the experiments by demonstrating the impact of model blindness and model blindness mitigation technique on performance. The experiments implemented a simulated route recommender system as an MDSS with a true data-generating model (unobservable world model). In Experiment 1, the true model generating the recommended routes along with additional non-recommended routes and the associated attribute information was misspecified to different levels to impose model blindness on MDSS users. In Experiment 2, the same route-recommender system was employed with a mitigation technique to overcome the impact of model-misspecifications on decision quality. Overall, the results of both experiments provide little support for performance degradation due to model blindness imposed by misspecified systems. The behavior captured in Experiments 1 and 2 showed minimal sensitivity to the different misspecified

statistical environments participants operated within. There was strong evidence of the impact of recommended alternatives and participants' reliance or deviation from them between conditions. The XAI intervention provided valuable insights into how participants adjusted their decision-making to account for bias in the system and deviated from choosing the model-recommended alternatives. The participants' decision strategies revealed that they could understand model limitations from feedback or explanations and adapt their strategy accordingly to account for misspecifications in the model. The results provide strong support for evaluating the role of decision strategies in the model blindness confluence model. These results help establish a need for carefully evaluating model blindness during the development, implementation, and usage stages of MDSS.

CHAPTER 1. INTRODUCTION

With the increasing popularity of Artificial Intelligence (AI) techniques, like machine learning, because of its proven ability to model a large amount of complex data and biological phenomenon, big tech companies, data scientists, medical professionals, and almost every large workplace is investing more time and funding to deploy these technologies (Doyen & Dadario, 2022). However, deploying these technologies without proper evaluation in complex socio-technical environments can have huge consequences on the end users and the organization. People are often faced with decisions in which they have to choose from the recommended alternatives provided by algorithms such as Amazon's product or Netflix's content recommender systems. Similar recommender systems, commonly known as Model-based Decision Support Systems (MDSS) (Power & Sharda, 2007), are often used to aid professionals' decision-making in many high-stakes, high-consequence domains such as aeronautics, emergency management, military command and control, healthcare, nuclear plant operations, intelligence analysis, and maritime operations. MDSS often uses task and operator models to provide the operator with cues, attributes, or decision alternatives to aid decision-making. MDSS are ubiquitous in many professional domains of high consequence. Given the prevalence of MDSS, it is surprising that work in model evaluation and human performance has not been leveraged to its full potential to evaluate these tools' characteristics and quality.

1.1 Thesis Motivation/Problem Statement

Analogous to how Google filters search results based on user search history and how social media websites (e.g., Facebook and Twitter) envelop their users into an information

bubble using past click-data to present customized information (Eady et al., 2019). Professional operators like intelligence analysts, pilots, physicians, etc., can be put into an accessible cue bubble via MDSS (Eady et al., 2019; Lawrence et al., 2018; van Leeuwen et al., 2021). This happens because models are simplifications, can be misspecified, and have errors (Marakas, 2003). The impoverished decision-making of the user due to the adoption, use, and limitations of model(s) is defined as **Model Blindness** (Parmar et al., 2021). There is a need for acknowledgment and empirical investigation of the issue of model blindness before implementing MDSS in a high-consequence decision-making task. This is important because users of MDSS might tend to lose perspective regarding model limitations after adopting and using decision support systems, which can lead to subsequent impoverished decision-making. The same term "model blindness" was coined in a similar context by an economist, Gittins (2012), to draw attention to simplifications of economic theories and rational models, which usually leave out important elements like irrational decision makers. This reduction of all complications from economic theories and models makes economists prone to the risk of an occupational hazard called "model blindness" (Gittins, 2012).

The consequences of errorful models are acknowledged by the US Department of Defense's AI ethical principles (Figure 1) adopted recently by the department for the design, development, deployment, and use of any AI capabilities by the DoD (Board, 2019). Figure 1 shows the DoD's five major principles that aim to make AI systems more ethical and useful, and all of them indirectly reach the concept of reducing model blindness or making users or designers aware of potential model blindness. The issues identified by DoD are not unique to the Defense community; recently, Doyen and Dadario (2022) also

identified similar challenges to AI implementations in healthcare and identified 12 plagues of AI (Figure 2) that are affecting both deployment and adoption of AI in healthcare. O'Neil (2016) also referred to the algorithms implemented in all domains of our lives, like credit scoring and employee evaluation algorithms, as "Weapons of math destruction" due to their hidden data variables and unreliable recommendations. Hence, we can see constant criticisms of these AI/model-based systems due to their limitations and the consequences they can have on people. The literature indicates a dire need for more formal investigations of these model-based systems in controlled lab-based and naturalistic decision-making environments before any real-world implementations.

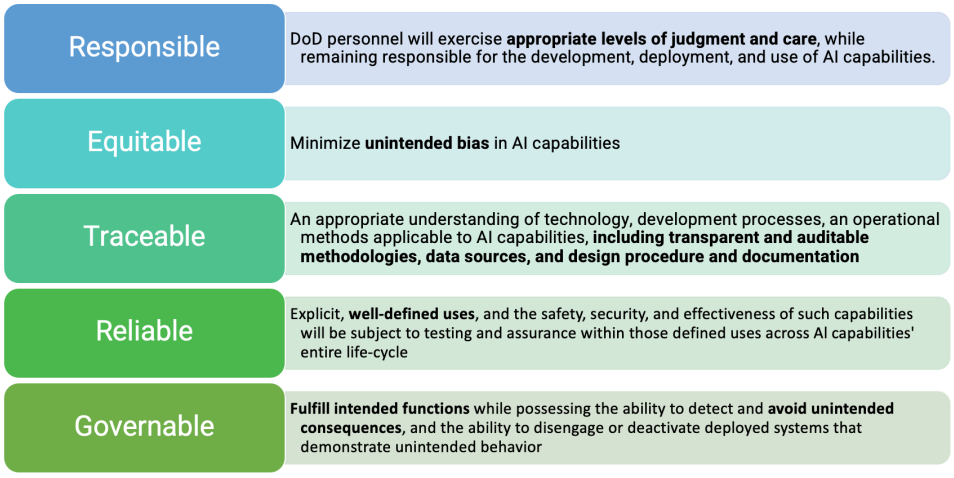

Figure 1. DoD's principles for responsible Artificial Intelligence (adapted from Board (2019))

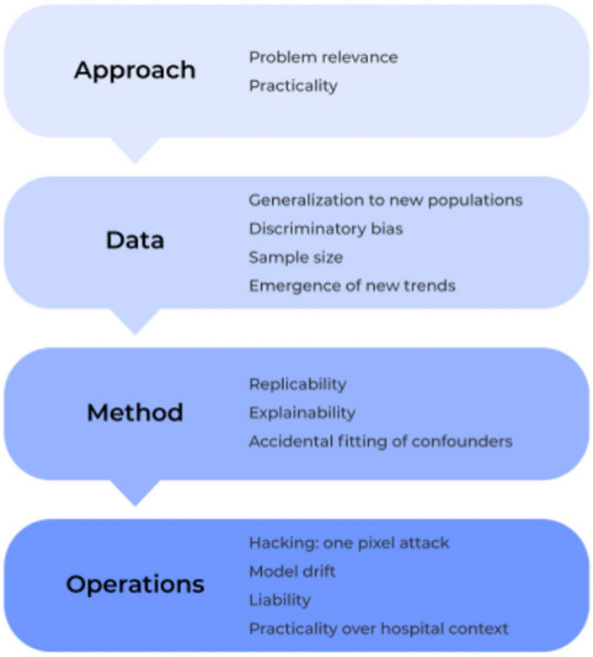

Figure 2. 12 plagues of AI in healthcare affecting the deployment process. Image taken from Doyen and Dadario (2022).

1.2 Research Questions

As recommended by DoD's principles, the results of this dissertation experiments can contribute toward making MDSS more traceable, reliable, governable, responsible, and equitable. This dissertation empirically evaluates the confluence model (discussed in detail in Chapter 2) of operator performance under model blindness proposed by Parmar et al. (2021) to understand the consequences of model blindness on human performance and explore ways to mitigate those consequences. The studies conducted to achieve this goal were designed to answer the following broad research questions (RQs):

RQ1: How does the model misspecifications in an MDSS impose model blindness on users? How does that manifest in their decision-making, performance, confidence, and trust in the MDSS (main effect of model misspecification)?

RQ2: How does presenting model-recommended alternatives with additional decision alternatives for decision-making task help or hurt a decision-maker's choice selection (main effect of recommender)? Does it interact with the level of misspecification in the model producing those alternatives (Interaction between recommender and misspecification levels)?

RQ3: How can misspecified models bias the human decision-making process and decision-strategy selection? When do users detect issues with the recommended alternatives presented by an MDSS? Do they deviate from MDSS recommendations or choose to continue relying upon them?

RQ4: How do decision makers optimize and adjust their strategies when there is bias or misspecification in MDSS that can be detected from available feedback? Do they adjust their strategy to overcome bias and match true world feedback?

RQ5: Does presenting natural language explanations help calibrate decision-makers to the capabilities and limitations of an MDSS and consequently improve performance? Does it mitigate the consequences of model blindness to some extent?

1.3 Study Summary

To test the research questions presented above, I conducted a series of two experiments by systematically controlling for misspecification levels in the MDSS model, recommended alternatives' availability, and explanation availability. Both experiments implemented a simulated route recommender system as an MDSS with a true data-generating model (unobservable world model). The MDSS presented seven routes, each

with six associated decision attributes to the participant, and participants were given the explicit goal of picking the best route. Chapter 3 discusses the experiments in detail, including how the MDSS was simulated.

In Experiment 1, I manipulated the MDSS model's misspecification level (no misspecification, soft misspecification, and hard misspecification) and the presence of MDSS-preferred recommended routes (three preferred routes vs. no preferred routes) as between-subjects manipulation. As shown in Figure 3, the true data generating model presenting the seven routes and the associated attribute information is misspecified at two levels (soft misspecification and hard misspecification). Random noise is added in all models used in the experiment as there is never a perfectly accurate model available in uncertain and high-consequence decision-making environments. Experiment 1 tested how direct manipulations of model blindness imposed by MDSS affect the decision-making process and performance. In Experiment 2, the same route-recommender system was employed along with a mitigation technique to overcome the impact of model-misspecifications (soft and hard) on decision quality. The mitigation technique provided participants with a natural language explanation (explainable-AI intervention) of misspecification in the model as a warning message before the decision task. The explanation in Experiment 2 helped investigate whether the consequences of model blindness can be mitigated by making users aware of the shortcomings of the model that are causing the MDSS to present lower-quality route alternatives.

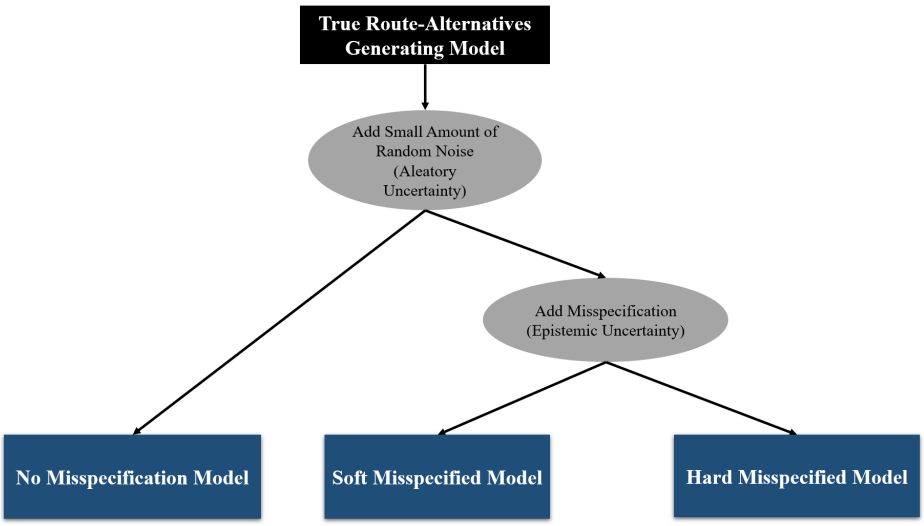

Figure 3. Models used in the route recommender system developed for experiments

Overall, the results of both experiments help address some of the research questions mentioned above. It provides little support for performance degradation due to model blindness imposed by misspecified systems. The behavior captured in Experiments 1 and 2 showed minimal sensitivity to the different misspecified statistical environments participants operated within. There was strong evidence of the impact of recommended alternatives and participants' reliance or deviation from them between conditions. The explainable-AI intervention provided valuable insights into how participants adjusted their decision-making to account for bias in the system and deviated from choosing the model-recommended alternatives. The participants' decision strategies revealed that they could understand model limitations from feedback or explanations and adjust their strategy accordingly to account for misspecifications in the model. The results provide strong support for evaluating the role of decision strategies in the model blindness confluence model. The confluence model will be discussed in detail in the next section while reviewing background research on this dissertation topic.

CHAPTER 2. BACKGROUND LITERATURE

MDSS are often implemented to make multi-objective decisions in domains with high uncertainty, high information overload, time pressure, and dynamic changes in the task. MDSS usually provides multiple (or single) courses of action recommendations with or without the associated information cues. The widescale implementation of MDSS in one form or the other indicates the enormous benefits of using MDSS. MDSS type DSS are widespread because (Beemer & Gregg, 2008; Marakas, 2003; Turban et al., 2001): (1) It decreases decision-making time, (2) Enhances the problem-solving and decision-making process, (3) Improves both decision making process and quality of decision, (4) It provides the ability to solve complex problems beyond human reach due to amount of data or processing needs to be required.

2.1 Prevalence of MDSS in High-Consequence Decision-Making Tasks

MDSS are prevalent in maritime operations and often involve generating multiple probabilistic courses of action for the operator. For example, TMPLAR (Tool for Multi-Objective Planning and Asset Routing) is a DSS developed to address the asset routing problem for the Naval and commercial shipping (Avvari et al., 2018). The tool provides recommended paths (see Figure 4) comprising waypoints and associated arrival and departure times, asset speed, and bearing. The recommended paths and associated schedules are optimized based on several objectives (decision attributes) like voyage time, distance, fuel efficiency, asset vehicle limits, navigator-specified deadlines, etc. Other tasks where TMPLAR-like tools are being frequently implemented are counter-smuggling operations (COAST-Courses of Action Simulated Tool) and dynamic autonomous aerial systems' operations under uncertainty (SCOUT- Supervisory Control Operations User Testbed) (Mishra et al., 2017).

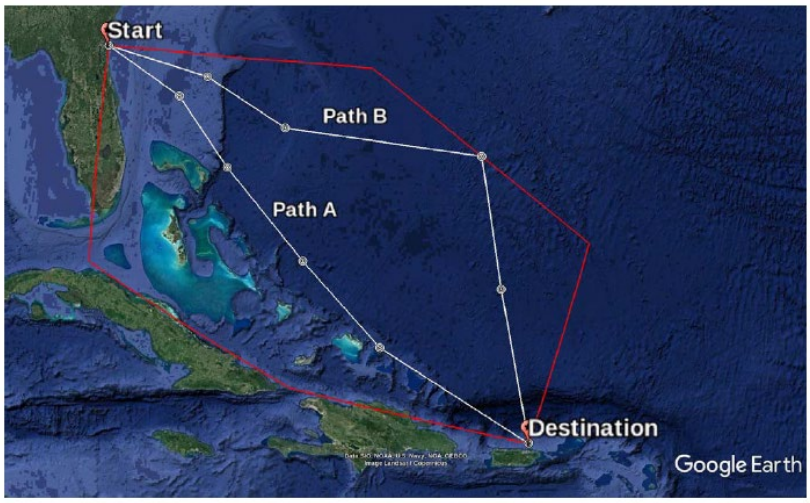

Figure 4. Path A and B are two paths recommended by TMPLAR's algorithm with the wait times schedule for each waypoint via which Hurricane Joaquin could be avoided. Image taken from (Avvari et al., 2018).

In healthcare, MDSS are implemented in the form of clinical decision support systems (CDSS), which primarily provide information or suggestions to the physicians at point-of-care to combine their knowledge with support from the CDSS (Sutton et al., 2020). CDSS are typically integrated with Electronic Health Records (EHRs) or Computerized Provider Order Entry (CPOE) (Sutton et al., 2020). A CDSS is usually used to provide patient-specific assessment or recommendations to a physician deciding on a diagnosis or treatment (Sim et al., 2001). AI-based medical services are also becoming common for tasks like automated evaluation of X-rays in radiology, skin cancer detection from patient-uploaded images in a mobile app, telehealth services, and so on (Cadario et al., 2021).

Although unnoticed, model-based systems exist in almost all aspects of everyday life, including low and high-consequential decisions. Many do not provide information cues (attributes) or alternatives but give only one final decision (presumed best by the algorithm). Especially in the healthcare domain, there is a lot of pushback and a lack of

trust by both patients and care providers concerning the use of automatic recommenders or AI-based systems, particularly when such systems seek to replace expert physicians (Longoni et al., 2019; Promberger & Baron, 2006). Algorithmic Decision-Making Systems (ADMS) (Rebitschek et al., 2021), like employee-selection aids for making hiring decisions, are also becoming increasingly popular and face a lot of distrust issues (Diab et al., 2011).

Many recommender systems implement algorithms to reduce choice overload and present the decision-maker with a reduced set of options to choose from (Jameson et al., 2015). Some recommender systems provide only one recommendation with or without explaining why the specific recommendation was provided (Jameson et al., 2015). Jameson et al. (2015) emphasizes the important distinction between models of choice (providing one single action choice or autonomous systems) and models of choice support (providing information cues or multiple action alternatives) in the context of these systems. This dissertation focuses primarily on models of choice support, but models of choice will face more dire consequences of model blindness.

2.2 Model Blindness using MDSS

Figure 5 presents a simplified architecture of an MDSS that mimics the framework of the route-recommender system used in this dissertation. As shown in the framework, there are five main components of an MDSS: database, decision support model, recommender model, user interface, and decision-maker. Although the limitations in model components (both decision support and recommender) of MDSS are primary sources of model blindness, the amount of model blindness imposed on users and its subsequent consequences can be exacerbated by limitations in all other components of MDSS as well. I explicitly focus on probabilistic (non-deterministic) models implemented in uncertain

domains. Model input can be affected by the data quality in the database, and model output can be affected by how information is presented on the user interface and how the human decision-maker processes it. Below are some of the limitations that each component of MDSS can suffer from, consequently imposing a state of model blindness on the user.

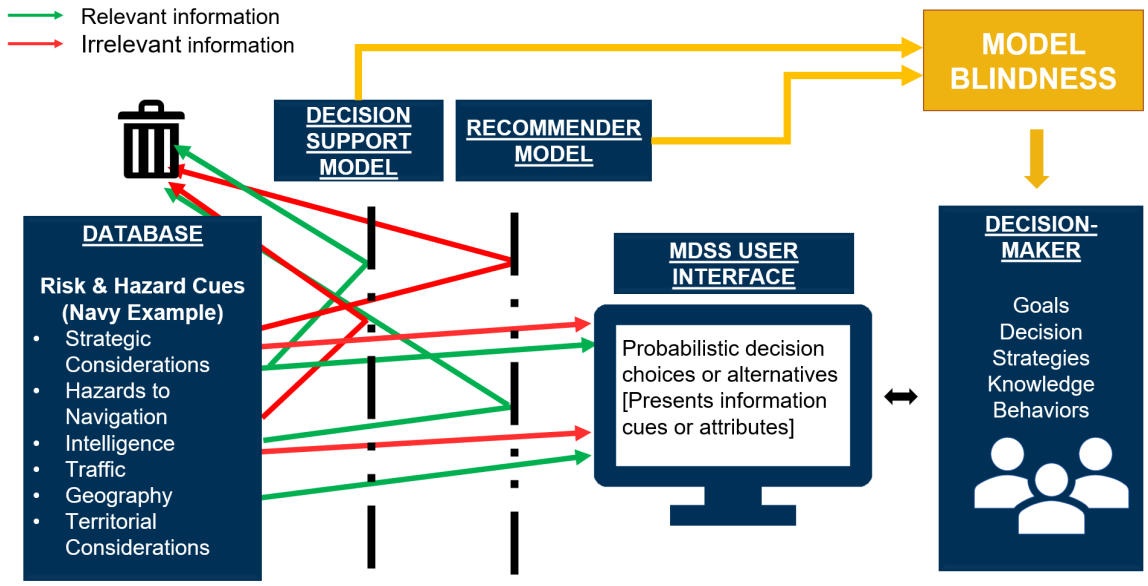

Figure 5. MDSS architecture with risk and hazard cue examples from naval ship navigation tasks

- A ***database(s)*** contains all possible information necessary for the decision-making task. It can suffer from errors, selection bias, and missing information (Baeza-Yates, 2016). E.g., In 2014, Twitter launched a bot, "Tay," whose tweets became sexist and racist within 24 hours because it learned from data gathered using biased user tweets (Baeza-Yates, 2016).
- A ***decision support model*** of MDSS uses algorithms to analyze information from the database to identify a few relevant alternatives (or relevant information) identified best by the model for a decision-making task. Hence, the decision support model imposes a filter that only lets some relevant information pass through it. This model can exhibit bias and fail to operationalize context, tasks, and operator variables appropriately,

limiting their ability to discriminate the relevancy of the information necessary to support decision-making, hence imposing model blindness on users.

- A subsequent ***recommender model*** imposes another filter on the alternatives identified as best by the decision support model. A recommender algorithm usually uses ranking-based algorithms to determine which alternatives are most relevant (preferred choice set) for decision-making. Hence, further reducing the choice set for users. This model is also prone to errors and misspecifications, leading to model blindness.
- Subsequently, the output of these model filters again goes through the visualization process imposed by the ***user interface***. The whole data visualization field attempts to help with aspects of presenting data before the data reaches a visual interface. The user interface can suffer from limitations due to information presentation and interaction bias (Baeza-Yates, 2016), cue accessibility, and saliency (Payne, 1980; Wickens et al., 2015). E.g., using loud sounds, bright flashing lights, highlighting some information cues, etc. These can subsequently impose another interface-imposed filter on the model output through which users receive the information.
- Finally, the ***decision-makers*** will also modify the information while processing information presented on the interface to make a decision based on their task objectives, knowledge, strategy choices, etc. The decision-makers come with their own biases (Mosier & Fischer, 2010), like automation misuse or disuse, over-trust or under-trust, and a tradeoff between goals; these add to the challenges faced by MDSS. E.g., blind compliance (Mosier & Fischer, 2010) by the user of MDSS can lead to less active information search and situation assessment or focusing on salient alternatives rather than evaluating all available alternatives.

2.2.1 Reasons behind MDSS component limitations

The presence of uncertainty in these complex, probabilistic, and multi-objective decision-making environments is the primary cause of these MDSS limitations, as

presenting completely accurate (deterministic) information is impossible. Moreover, because there is a tradeoff between avoiding information overload and providing all relevant information to operators—one can never provide all the available information (Woods et al., 2002)—model blindness results from a tradeoff that cannot be entirely eliminated. There are two sources of uncertainty in the multi-objective decision-making tasks that I am focusing on: epistemic and aleatory (Cerutti et al., 2022; Fox & Ülkümen, 2011; Hora, 1996). Aleatory uncertainty refers to the inherent irreducible uncertainty due to probabilistic variability (inherently random effect, e.g., flipping a fair coin). Epistemic uncertainty is the scientific uncertainty in the model of the process due to limited data and knowledge. Both sources of uncertainty will be operating in the route recommender system simulated for this dissertation. Going back to Figure 2, the added random noise to the true data generating model and the probabilistic nature of the task I model is clearly aleatory. Epistemic is more akin to model misspecification (cues misweighted or missing) due to lack of knowledge. I think one acknowledges epistemic uncertainty whenever one use a model—all models are simplifications (Box & Luceno, 1997, p. 6).

The filters arising in the MDSS in Figure 5 due to these uncertainties operating in the decision-making environment can be systematic (or intentional) or emergent (unintentional). These filters can arise from issues ranging from clear and known design or modeling choices that intentionally leave some information out of the algorithm, like social media information bubbles. These filters can also emerge due to constraints in the decision environment, like noisy information, incomplete knowledge, and tradeoffs. Shah and Oppenheimer's (2008) effort reduction principle posits that decision-makers tend to use the information cues that are most easily accessible (easy to retrieve from memory or made readily available by other means like DSS) in their decision-making. Therefore, the less accurately MDSS identifies relevant information and the extent to which access to irrelevant information is enhanced, the more likely it is that decision-making performance

will degrade. As domains become more complex and the reliance on models to aid decision-making increases, the implications of model blindness will become increasingly important.

2.3 Confluence Model of Operator Performance Under Model Blindness

In previous work (Parmar et al., 2021), we proposed a confluence model of operator performance under model blindness and demonstrated the impact of model blindness on performance via simulations. The proposed confluence model (Figure 6) posits that performance degradation due to model blindness imposed on the decision-maker depends on the match or mismatch between the interaction of three components- model, decision context, and human decision strategy. These components can consequently lead to model-limited, data-limited, or strategy-limited performance. This taxonomy is inspired by Norman and Bobrow's (1975) theory of data-limited vs. resource-limited processing, where resource-limited processing occurs when an increase in the amount of processing resources allocated to the task leads to improvement in performance. Contrary to this, in data-limited processing, performance is only influenced by the quality of data available, and no amount of additional processing resources can enhance performance in a task. Performance degradation and decision quality are often operationalized in terms of proportion correct, sensitivity, choosing Pareto-optimal alternatives, and expected value or utility in decision-theory literature (Payne et al., 1993).

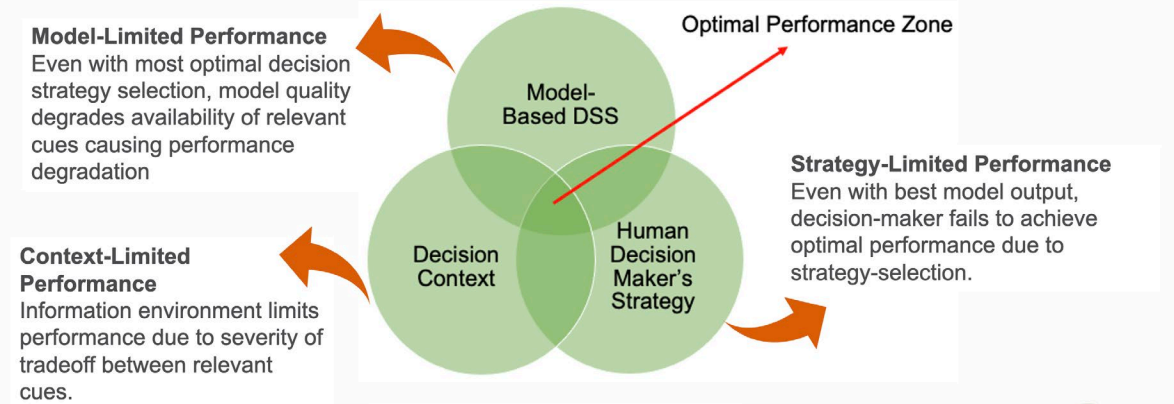

Figure 6. Confluence model of operator performance under model blindness

2.3.1 Model-limited performance

Model-limited performance manifests when the degradation of performance occurs because of the limitations of the model. One way to operationalize model-limited performance is when the number of relevant cues provided by the MDSS is less than that which can be utilized via the decision maker's decision strategy. Another way to operationalize model-limited performance is when misspecifications in the model degrade the quality of the alternative set presented to the decision-maker and consequently impact the decision quality.

2.3.2 Strategy-limited performance

Decision-maker's strategy selection is a mode by which an individual evaluates alternatives and associated attributes and makes a final choice out of multiple presented options. People's decision strategies can broadly be classified as compensatory or non-compensatory. Compensatory processes involve making tradeoffs between attributes (e.g., weighing differentially, adding pros and cons) (Kurz-Milcke & Gigerenzer, 2007). In

contrast, non-compensatory processes make no trade-offs (e.g., heuristic strategies like Take The Best) (Kurz-Milcke & Gigerenzer, 2007).

Strategy-limited performance manifests when performance degradation occurs because of the limitations in the decision maker's strategy selection. Even with the best model output (all relevant cues or top alternative set), the decision-maker fails to achieve optimal performance. For this dissertation, I operationalize strategy-limited performance as the number of information cues provided by the model exceeds the utilization of the decision maker's strategy, particularly under time pressure or attention (dual-task) demands (Payne et al., 1993).

2.3.3 Context-limited performance

Context-limited performance manifests when the information environment limits performance. One way to think about how the information environment can limit performance is by the severity of the tradeoffs between attributes. When an environment is friendly (positive inter-attribute correlations), the decision-maker can maximize all attribute dimensions (Shanteau & Thomas, 2000). However, the decision-maker cannot maximize all attributes when the environment is unfriendly due to tradeoffs (negative inter-attribute correlations)—the decision-maker must give up value on one attribute dimension to gain value on others. When decision-makers implement less-than-optimal decision strategies, performance degradation tends to worsen as the inter-attribute tradeoffs increase in severity (Payne et al., 1993; Shanteau & Thomas, 2000). The experiments in this dissertation only implemented an unfriendly decision environment due to the unfriendly

environment being a more common problem for MDSS in complex multi-objective decisions.

2.4 Framework to Empirically Evaluate the Confluence Model

I propose a framework (Figure 7) to evaluate the implications of model blindness for the three categories of performance degradation shown in Figure 6. In the framework, the MDSS comprises a filter (the detailed version shown in Figure 5) by which the decision support and recommender model implicitly filter out some cues while making others available to the decision-maker. The accuracy of the model filters will impact the model blindness imposed on the operator's performance. Note that the extent to which the negative consequences of an inaccurate model filter impact performance will depend on the decision context and the selected strategy. The decision strategy lens adapted from Brunswik (1952) will determine how different cues presented by the model will converge to final decision on the human side via their strategy selection, leading to strategy-limited performance. Some strategies listed inside the lens use fewer cues than others or weigh cues disproportionately, causing a potential for strategy-limited performance in some decision contexts. The decision context (friendly vs. unfriendly) will also impact how the final decision will be model-limited. Decision context will act as a moderator for the model's impact on performance leading to operator performance being more or less likely to be model limited in unfriendly and friendly environments, respectively.

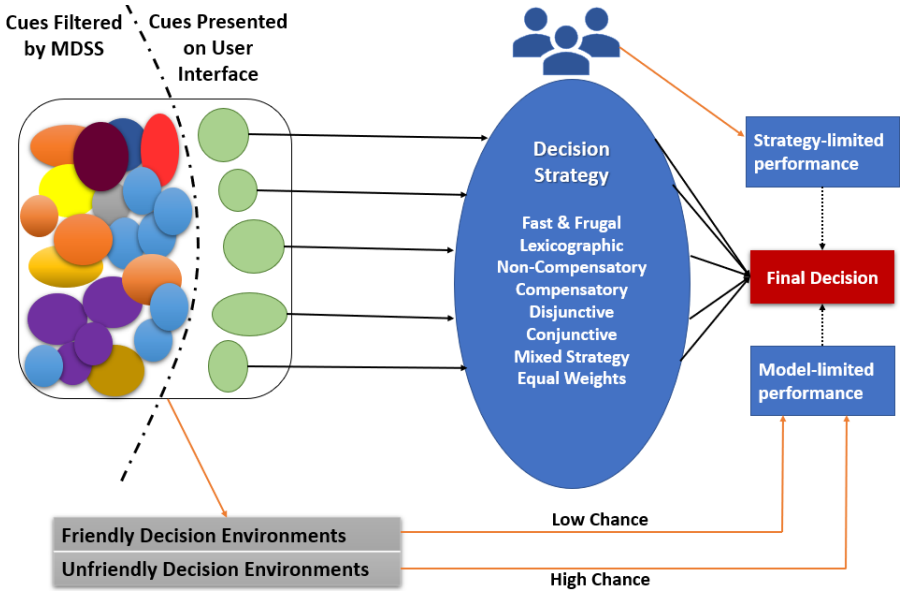

Figure 7. Framework for model blindness's impact on performance showing how the three components of the confluence model interact with each other

Researchers can leverage the framework to address several essential questions empirically. What type of DSS can minimize the impact of model blindness? What is the minimum permissible level of model blindness that can still result in optimal performance? Which decision strategies are most robust to the model blindness imposed by a particular system or aid in a specific decision context? What levels of unfriendliness in the environment cause substantial performance degradation? What mitigation approach works best to reduce performance degradation imposed by MDSS? By manipulating various independent variables in an experimental setting as either between-subjects or within-subjects factors to measure performance degradation, these important questions, as well as others, can be investigated empirically.

For example, the level or amount of model blindness imposed by a tool can be manipulated via intentionally presenting irrelevant cues along with relevant cues, filtering out some relevant cues, or withholding relevant cues. Similarly, the level of MDSS can be manipulated in ways ranging from providing only information cues to providing multiple possible ranked probabilistic decision alternatives to evaluate the impact of model blindness imposed by different levels of automation for DSS on human performance. Similarly, environment friendliness or unfriendliness can also be manipulated. Also, different mitigation strategies for model blindness can be tested similarly.

This dissertation aims to evaluate the proposed framework empirically by manipulating the level of model blindness in an MDSS via misspecifying the model that will affect the quality of the alternative set along with the quality of information attributes presented to the participant (Experiment 1). The participants will be recommended to follow one decision strategy in an unfriendly decision context. The next sub-section presents some simulation work done to evaluate the framework and motivate the proposed experiments for this dissertation.

2.5 Simulations to demonstrate how performance degradation can manifest via model blindness

This section presents a series of simulations to demonstrate how performance degradation can manifest via model blindness. The general simulation methodology was an abstraction of a naval fleet-movement task using custom code developed in Wolfram Mathematica (Parmar et al., 2021). The steps in the simulation are illustrated in Figure 8: map generation, defining areas of hazards (areas to avoid) and objectives (areas to complete

mission objectives), including their size and overlap, application of values and probabilities to the hazard and objective areas, route generation (many 1000s were generated), simulated mission planners that down-selected the routes, and decision making (route selection) via simulated commanders. The disjunctive probability that the route alternatives led to successful mission completion (avoided hazards and completed objectives) defined their objective utilities. Also, different MDSSs were simulated by the amount of error they exhibited. The simulated commanders utilized the different decision strategies (see Table 1) that had access only to cues (route attributes) provided by the hypothetical MDSS. Specifically, the route alternatives provided to the simulated commanders were defined by six attributes or cues reflecting the values and probabilities of hazards (e.g., air and sea threats, etc.) and objectives (e.g., surveillance and reconnaissance missions) derived from the simulated map. We implemented the proportion of utility loss as the performance measure (see Equation 1), where higher numbers translate to worse performance.

$$Proportion\ Utility\ Loss = \frac{Utility_{Best\ Alternative} - Utility_{Selected\ Alternative}}{Utility_{Best\ Alternative}} \quad (1)$$

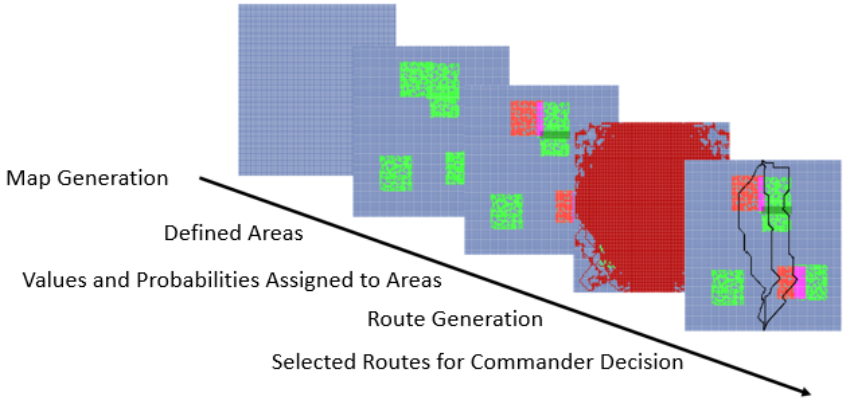

Figure 8. Illustration of the general simulation methodology of the naval fleet-movement task.

Table 1. Decision strategies used in the naval-fleet movement task simulations

Decision strategy	Description
Weighted Additive (WADD)	Apply true weights to cues and select the option with highest sum
Maximax	Selects the option that maximizes the probability of best-case outcome
Equal Weights	Weight each attribute 1/n and select option with the highest sum
Maximin	Picks the option that minimizes the probability of worst-case outcome
Conjunctive	Options must exceed some minimum threshold for each attribute
Disjunctive	Options must exceed some minimum threshold for only one attribute
Lexicographic	Assess options via attributes based on validity order; select an option based on the first cue that discriminates

2.5.1 Model Blindness due to information cues' quality

In the first demonstration (Figure 9), we simulated hypothetical MDSS of varying levels of model blindness where only a subset (3 most relevant vs. 3 least relevant) cues were available to the simulated commander (we provide the performance under the complete set of 6 cues for comparison). Unsurprisingly, the simulation results indicated a larger utility loss when the hypothetical MDSS presented the least-relevant cues to the simulated commander. Note also that the simulation demonstrates that the strategy used by the operator can contribute to the performance decrement or the robustness of the decision-making against utility loss due to model blindness. For example, the Maximax strategy shows robustness to the level of blindness in this decision context, although the robustness comes at the cost of suffering utility loss under all model blindness conditions. Note that if we had implemented a different utility function (e.g., survival probability), we would expect Maximin to have exhibited robustness to catastrophic failure (Ben-Haim, 2006).

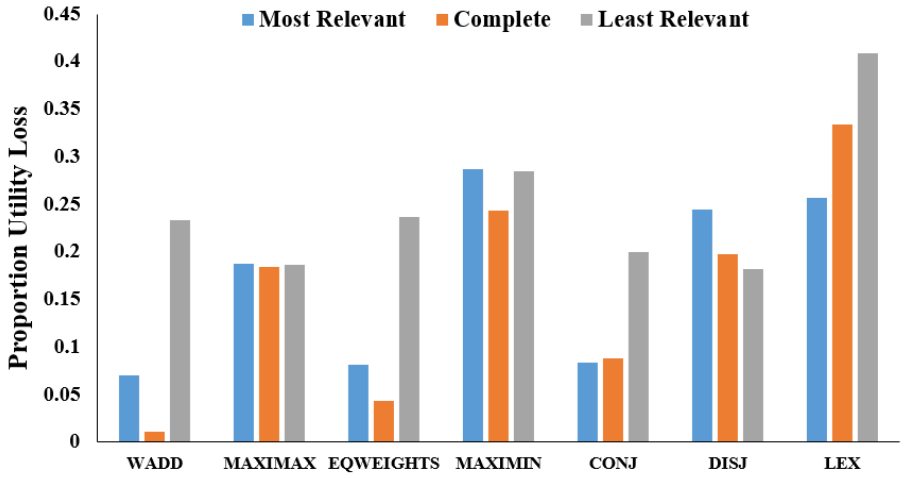

Figure 9. Utility loss under different levels of model blindness (information-quality) and decision strategies

2.5.2 Model Blindness due to decision choice alternatives

The following demonstration illustrates a situation where the hypothetical MDSS and the simulated mission planners provide alternatives to support the commander's decision. In this simulation, the MDSS presents Pareto-optimal (high-quality) alternatives, dominated (low-quality) alternatives, or alternatives via a random draw for the simulated commanders. The simulation results (Figure 10) indicate more utility loss when the hypothetical MDSS had the potential to select alternatives that were not Pareto-optimal. Again, this simulation demonstrates the limitation of the MDSS leading to more utility loss via model blindness from withholding higher-quality alternatives from the commander. It is important to note that the hypothetical MDSS is similar to many existing systems that can generate 1000s of alternatives for decision-makers to choose from, so context-sensitive rules are needed to filter out what are estimated to be lower-utility options. The models' ability to deliver high-quality alternatives given the decision context is necessary to avoid utility loss.

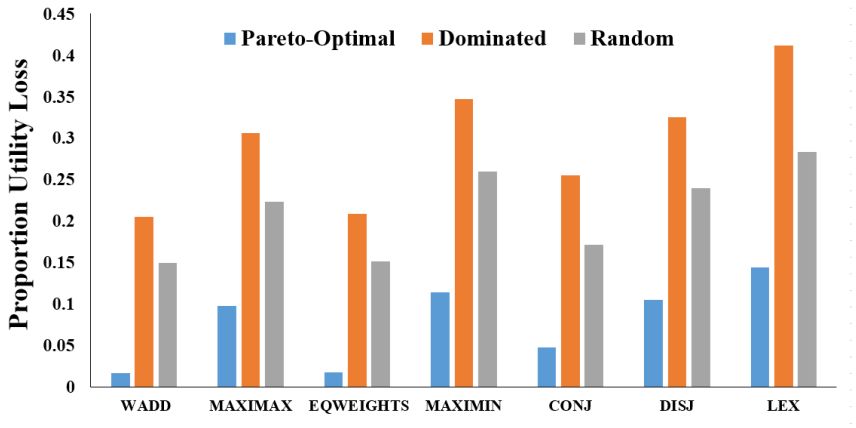

Figure 10. Utility loss under different levels of model blindness as a function of the quality of the alternatives provided by the MDSS (Pareto-optimal, Dominated, and Random) and decision strategies.

2.5.3 Other variables affecting performance

Task characteristics such as time pressure and cognitive load can affect the amount of information (number of cues) that operators can attend to and utilize. Both time pressure and cognitive load can be characterized as imposing a sort of blindness on the operator who cannot incorporate the complete set of MDSS-presented cues into their decision strategy. Specifically, we implemented time pressure and cognitive load by limiting each decision strategy to a random draw of only three (out of the six) presented cue values.

Figure 11 illustrates the results from a simulation where the operators' attentional access to the presented cue set via the MDSS is compromised due to time pressure or insufficient capacity (because of dual-task cognitive load). Although the compensatory decision strategies (CONJ, WADD, EQWEIGHTS) exhibit the least utility loss overall, they show a more considerable relative increase in utility loss under time pressure or cognitive load than the heuristic (non-compensatory) strategies.

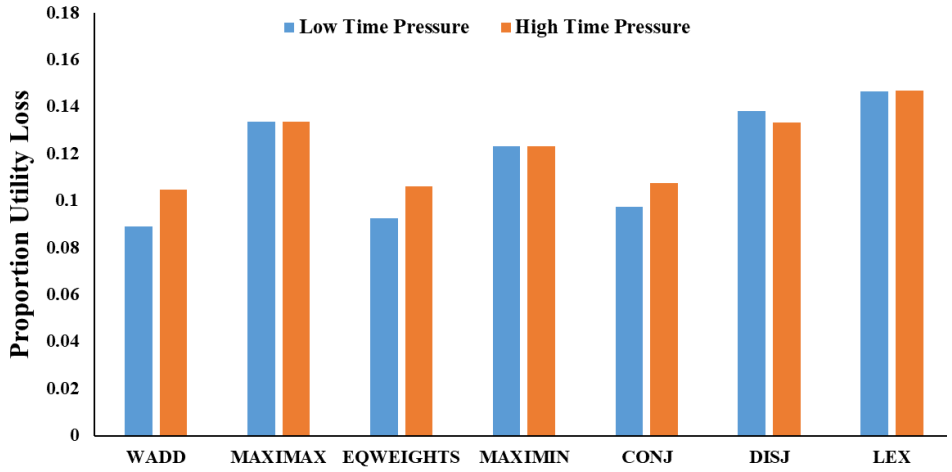

Figure 11. Utility loss results from varying levels of time pressure/cognitive load while using different decision strategies.

The simulation results presented in this section show how the quality of the alternative set, relevancy of information cues, and extrinsic factors like time pressure and workload are really important to prevent performance degradation for users using different decision strategies while processing information presented by the MDSS. Experiment 1 in this dissertation aimed to test the results found via these simulations in an empirical setting by manipulating both quality of the available alternative set and information cues associated with those alternatives via introducing misspecification (epistemic uncertainty) in the model of the MDSS.

2.6 Model Blindness Mitigation Strategies

The previous sub-sections in this chapter establish a need to acknowledge and evaluate the challenge of model blindness posed by MDSS. The next obvious step is the need to explore ways to mitigate the consequences of model blindness on users. Model blindness mitigation strategies can be developed by borrowing from the already growing research areas in human factors psychology and computer science of system observability (Woods

& Sarter, 1998), transparency, explainable AI (Páez, 2019), counterfactual reasoning capabilities, and so on. However, none of these areas address all three components of the proposed confluence model. Some recommendations for designers and developers of MDSS include (1) developing smart filtering models which are context-sensitive, (2) providing what-if capabilities (Heuer Jr & Pherson, 2010), (3) using ensemble models of tasks and users to reduce errors in the models' (Mangiameli et al., 2004), (4) providing decision influence diagrams (Papamichail & French, 2005), (5) communicating where the operators are in the trade space of information overload vs. providing all relevant information, (6) providing access to hidden information cues upon request from the user, (7) providing explanations and confidence in model recommendations, and (8) communicating to the operators about misspecifications or errors in the model. In general, there is a need to make users aware of the model blindness.

2.6.1 Simulations for Model Blindness Mitigation via Ensemble Modeling

The simulation results in Figure 12 show an example of a model blindness mitigation technique. The utility loss experienced as a result of model blindness can be at least partially mitigated via ensemble models, where the ensemble model in the simulated context was an aggregation of several "noisy" single models into a composite. The simulation is similar to the one presented in Figure 9, where a hypothetical MDSS presents a subset of cues to the operators to support their decision-making. The sigma at the bottom of the panels of Figure 12 indicates the level of error associated with a single model or a group (ensemble) of models providing the cues to the operators. The utility loss associated with model blindness is mitigated to the degree that the model(s) suffers lower error. Although the ensemble models in this simulation didn't replace missing or withheld cues,

the cues made available to the simulated operator via the MDSS were more accurate (via averaging across the ensemble models) when presented to the operators leading to lower utility loss. Once again, this simulation demonstrates that the performance decrement or the robustness of the decision-making against the utility loss imposed via model blindness varies greatly with respect to the decision strategy used by simulated decision-makers.

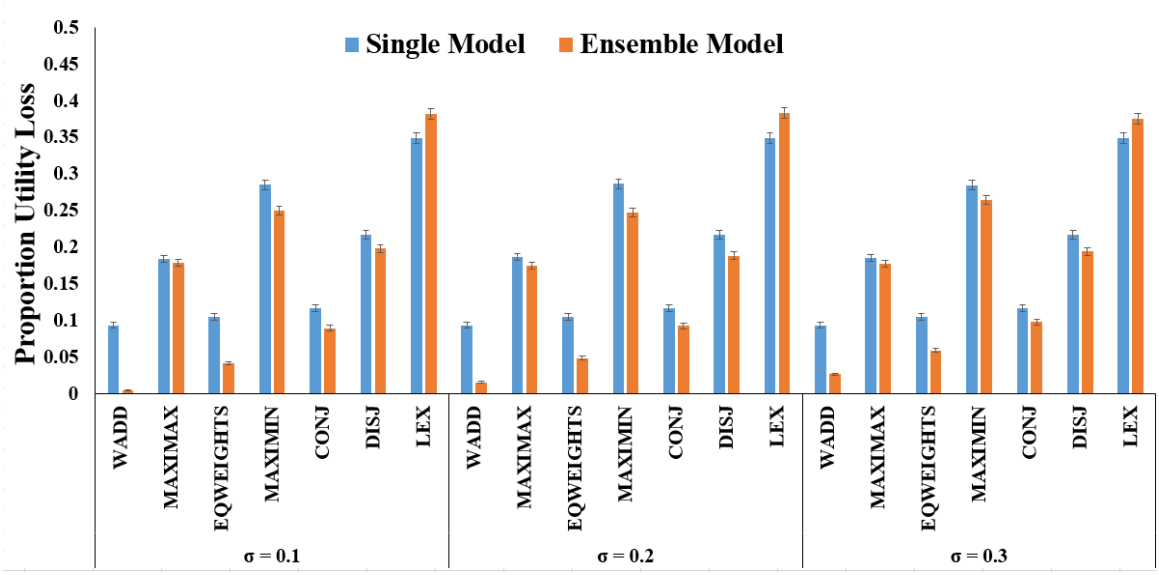

Figure 12. Utility loss of operators using different decision strategies with access to a single or ensemble suite of models with varying levels of error or uncertainty (higher values of sigma represent more model error)

2.6.2 Model Blindness Awareness

Similar to the concepts of Situation Awareness (SA) (Endsley, 1995) or Risk Situation Awareness (RSA) (Parmar & Thomas, 2020), there is a need for MDSS developers to design for operator's model blindness awareness. Being aware of model blindness can facilitate users to shift their decision strategies, impute missing information, spend more time evaluating the role of missing or irrelevant information, etc. There is a metacognitive aspect of knowing some information is missing or low quality vs. not

knowing at all. Designing the MDSS to shift its functioning modes is not enough to promote a shift of decision strategies or imputation of missing information by operators. Communicating where operators are in trade space can lead to user calibration by making the user aware of the level of blindness imposed by the system. In the context of adaptive automation, Woods and Cook (2006) emphasize that the adaptive capacity of a system should be known to the user. Most current adaptive systems make users learn through tangible experiences regarding how systems behave under disruptions and abnormal conditions, neglecting the role of active and constant user calibration to system capabilities. Similarly, in the case of model blindness, it is not possible for users to know what is misspecified in the MDSS. Although they might be able to understand what is incorrect from their experience with the DSS, like physicians and nurses learning to ignore irrelevant alerts given by Electronic Health Records due to desensitization to alerts (Kizzier-Carnahan et al., 2019); this calibration technique is not optimal.

Mitigating model blindness, in general, can help systems become more resilient. Resilient systems can promote cooperative cognition between humans and machines rather than merely building technologically advanced systems (Woods, 1993). Calibration of users to MDSS capabilities and limitations can prevent over-reliance or under-reliance on MDSS. Providing explanations for why information is presented or hidden can make users more aware of the system's capabilities and contribute to explainable AI (XAI) literature, which currently only focuses on explaining the model's outcomes.

2.6.3 Explainable AI

The underlying hypothesis behind building more transparent, explainable, or interpretable systems is providing users the capability to understand the system leading to better trust in intelligent systems (Chen et al., 2014; Mercado et al., 2016; Miller, 2019). Explainable Artificial Intelligence (XAI) is the area of research that facilitates different

types of understanding of the system. XAI researchers aim to build interpretative models, which are an accurate and reliable representation of black-box models used by AI agents, provide information about model limitations, and provide good comprehensibility for the user (Páez, 2019).

Providing complete transparency and understanding of a model is a challenge in uncertain and complex problems. Páez (2019) argues that explaining training sets, weights, and biases for complex and opaque machine learning (ML) models will not help understand the model's outcome. Because finding a causal pathway is not possible in models like stochastic (non-deterministic) models, deep neural networks, and so on. For such a model, the author emphasizes that the focus should be on understanding the model's decisions (or outcomes) rather than understanding the model itself. Researchers support this approach because it also helps provide explanations that cater to the user's level of expertise or understanding rather than the developer of the system. According to Páez (2019), the explanations do not require a truthful representation of these complex black-box models to make them understandable. These explanations' main goal is not to explain the functioning of the underlying model but provide valuable information for users and practitioners in an understandable way (Ehsan & Riedl, 2019). Diagrams, graphs, or maps can also be presented to enhance understanding without any explicit explanations (Páez, 2019). These explanations are generally used as a form of post-hoc interpretability in XAI literature to build confidence and develop understanding between humans and AI agents, especially in situations where AI fails or behaves unexpectedly (Ehsan & Riedl, 2019).

Ehsan and Riedl (2019) conducted a case study to show the benefits of using automated rationale generation to provide natural language explanations of an AI agent

playing a game involving a sequential decision-making task (past actions affect future actions). The study implemented a neural network to translate game state and action information into natural language explanations. The results identified a need for correct contextual data to make the automated rationales satisfying (adequate justification) for end-users. Users were also found to prefer detailed rationales that help build a complete mental model of users. Computer-based explanations to improve the trust and confidence levels of users are widely used in some form or the other (Dhaliwal & Benbasat, 1996; Papamichail & French, 2005). Dodson et al. (2013) also tested the efficiency of the system that generates conversational English language explanations for an academic advising system for completing courses to earn a degree.

I argue that methods in the XAI research, although capable of reducing model-limited performance, can potentially enhance the problem of model blindness instead of mitigating it. There are multiple reasons for this: (1) Adding a rationale generation component to an already blind model of an AI system can lead to overconfidence and overtrust rather than awareness of model blindness. (2) To explain one opaque model causing blindness, we are adding another one (explanation-generation model). Now users are double-blind – both due to misspecifications in the AI model and the XAI model of plausible rationales to reach one final reasoning, (3) These methods are mostly focused on model output. They are not doing a great job of helping the operator understand the conditions under which the model will perform well and when it is likely to fail.

Increasing the trust and reliance on AI agents is a primary goal for XAI research. Lack of user trust is considered one of the major challenges for ML models used in high-stakes environments of parole decisions or medical diagnosis (Páez, 2019). Trust in

automation is also a widely researched topic in the human factors literature and is implemented to improve operator reliance and calibration on automated systems. However, I argue that there can be unintended consequences for improving trust in the MDSS in the presence of model blindness. Explaining an MDSS's outcomes to the operator can be harmful if irrelevant information is not getting filtered out by the model. Presenting and explaining an irrelevant information cue or decision choice can instill a development of an inferior mental model by the decision-maker. It will worsen the impact of model blindness because the operator might already inherently believe that it has to be relevant in some form if the system presents something.

Experiment 2 of this dissertation mitigated model blindness imposed on users via MDSS model misspecifications. The mitigation technique used an XAI intervention aimed to create model blindness awareness instead of the traditional XAI approach of improving reliance and trust. Participants were provided with natural language explanations containing information about misspecification in the model of an MDSS. Results show how users adjust their decision-making, trust, reliance, and confidence in MDSS when they are made aware of the misspecifications. The next section discusses the methodology and results of both experiments in detail.

CHAPTER 3. METHOD AND RESULTS

The two experiments reported in this section utilize a simulated MDSS adapted to address different research questions posed previously. The experiments implemented a simulated model-based route recommender system (Figure 13) as an MDSS that participants used for a route planning decision task. The task required participants to find the best route to deliver critical shipments for COVID-19 patients from one geographical location to another via a route that involves a majority fleet movement and some ground shipping. Each decision-maker was presented with seven routes as decision alternatives. Each route was defined by utility values (ranging from 0 to 100) on six critical attributes to be evaluated for route-selection decisions. The higher the value of the decision attribute, the better utility it has, so participants were expected to maximize the value of all attributes for optimal performance. The six attributes were assigned arbitrary labels based on objectives important for routing problems— some of them adapted from Avvari et al. (2018) attributes for TMPLAR's asset routing problem and attributes used by Illingworth and Feigh (2021) for the disaster relief planning task. The attribute labels include time efficiency, fuel efficiency, obstacle avoidance (e.g., road closures, route deviations), weather hazard avoidance (flooded roads, hurricanes), en-route availability of extra supplies (vaccine shipments, medication shipments), and en-route ability to do humanitarian aid (deliver materials to people in need).

Figure 13. Route Recommender System interface for control groups in Experiment 1

3.1 Model-based Route Recommender System Development

The development of MDSS for this study involves (1) generating attribute values for all six attributes associated with all seven routes presented by the MDSS and (2) generating images with seven routes superimposed on a world map. Both steps were performed using two different custom codes in Wolfram Mathematica.

3.1.1 Model-Based Attributes Value Generation

For this study, I chose an equal weights model as the true data-generating (Figure 3) model for the MDSS. It was referred to as the true model because it reflected the true state of the world for the MDSS. An equal weights model requires that all attributes be given equal value to achieve an optimal decision (See Equation 2). The equal weights model made the decision-making task of appropriate complexity because the study implemented an unfriendly decision environment involving a negative inter-attribute correlation between three pairs of attributes (a total of six attributes), leading to increased sensitivity

to weights for an optimal decision due to tradeoffs. Participants also received instructions that strongly implied using an equal weights decision strategy to ensure they incorporated tradeoffs while making route-planning decisions (See APPENDIX A for detailed participant instructions).

$$Utility\ (Equal\ Weights\ Model) = \sum_{i=1}^{6} (Attribute\ Value \times \frac{1}{6}) \qquad (2)$$

The flow chart in Figure 14 lists the steps (from Mathematica code) involved in producing one trial (six attribute values for all seven routes) for MDSS. For each trial, many route alternatives (150 routes) were generated from which the true model found the seven best routes based on the highest ranked utility values (See Equation 2). The attribute values for all six attributes for 150 routes were generated by randomly drawing pairs from a copular binomial distribution with an average correlation of -0.7 between pairs of attributes (3 pairs). This negative correlation created unfriendly environments characterized by severe tradeoffs between decision attributes. The true model gave each attribute an equal weight of 1/6 (~0.16) and used the attribute values obtained from copular distribution to calculate the utility (See Equation 2) of all 150 routes. Following this, utility loss (see Equation 1) for each route was calculated with reference to the route with the highest utility. The top seven routes with the lowest utility loss values comprised the best alternative set that the true MDSS presented to the participants. However, as discussed before, models are always simplifications, and a true 100% accurate model doesn't exist in real-world scenarios when MDSS are implemented in uncertain decision-making tasks.

Due to this, the true model was not presented to the participants but was used to provide feedback and evaluate participant decisions compared to the true state of the world. The participants received seven routes (along with their decision attribute values) down-selected from 150 routes by either an accurate (no misspecification), soft, or hard misspecified model, as discussed earlier in Figure 3.

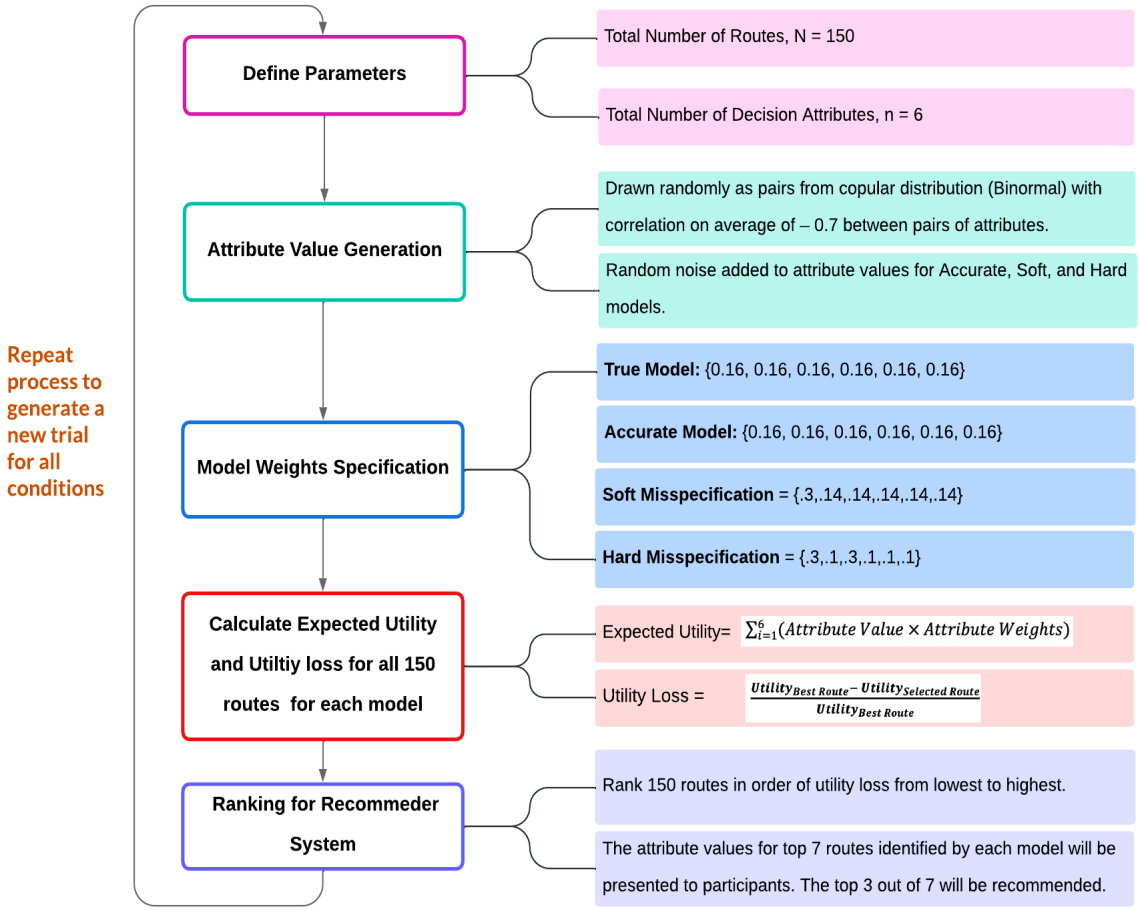

Figure 14. Steps to generate a set of 6 attribute values for the top 7 routes and top 3 recommended routes by an MDSS

Random noise (Normal Dist [0,0.015]) was added to the attribute values for all 150 routes as real-world data always has some noise level due to multiple sources of imperfections, e.g., errors in sensors collecting data leading to aleatory uncertainty. The

accurate model used the exact weights of 1/6 as the true model but used noised attributes to calculate utility and utility loss for each route and ranked them to identify the top seven routes and corresponding attribute values. Along with aleatory uncertainty, the misspecification models had epistemic uncertainty due to incorrectly weighted cues. The soft misspecified model was misspecified such that it overweighted the first attribute (2/6~almost double) and equally weighted the other five attributes such that the total weight sum is equal to one. As the name suggests, the hard misspecified model was misspecified to a greater extent than the soft. It overweighted the first and third attributes (2/6~ almost double) and equally weighted the other four such that the total weight sum was equal to one. Table 2 shows the attribute weights of each model.

Table 2 Attribute weights for each model (misspecification in bold)

	Attribute 1 Time	Attribute 2 Fuel	Attribute 3 Obstacles	Attribute 4 Extra Supplies	Attribute 5 Hazards	Attribute 6 Humanitarian Aid
True Model	0.167	0.167	0.167	0.167	0.167	0.167
Accurate Model	0.167	0.167	0.167	0.167	0.167	0.167
Soft Misspecified Model	**0.3**	0.14	0.14	0.14	0.14	0.14
Hard Misspecified Model	**0.3**	0.1	**0.3**	0.1	0.1	0.1

The utility and utility loss for the top seven routes and attribute values were different in all four models because of the different weights (Table 2) the model gave while

producing the seven alternatives. Hence, the set of seven alternative routes differs between each model. Table 3 shows the output ranks (based on the true model) for one trial for each model's top seven route alternatives. For example, the table shows that the 55th ranked route by the true model is identified as the third best route by the hard model, hence showing the implication of model misspecification on the quality of alternatives presented by the model.

Table 3. Example trial: True ranks (out of 150 routes) of route alternatives ranked as top 7 by each model

Model-based route ranks	True ranks of routes out of 150 routes used to find 7 top routes by each model			
	True Model	Accurate Model	Soft Misspecified Model	Hard Misspecified Model
1st best route	1	1	11	21
2nd best route	2	2	21	27
3rd best route	3	3	12	55
4th best route	4	5	3	18
5th best route	5	4	26	41
6th best route	6	8	4	26
7th best route	7	7	27	11

The whole process in Figure 14 was repeated forty times to generate forty trials with seven alternatives and associated attribute values from all four models. Table 4 shows the descriptive statistics (mean true rank and standard deviation) for the top seven routes

generated by all three models for all forty trials used in the experiment. The first, second, and third best routes became the model-recommended set of routes for the participants and were presented in a different color for the experimental conditions. The details of the experimental conditions will be discussed in the next section.

Table 4. Descriptive statistics for true route ranks (out of 150) of top seven routes by all models for all 40 trials used in the experiments

Model-based Rank of Routes	Accurate Model		Soft Misspecified Model		Hard Misspecified Model	
	Mean Rank	SD	Mean Rank	SD	Mean Rank	SD
1st best route	1.10	.30	3.10	3.47	8.78	10.87
2nd best route	2.05	.68	4.58	4.01	13.23	12.99
3rd best route	3.24	.53	7.10	6.20	20.93	24.46
4th best route	4.24	1.02	8.38	6.11	20.87	21.79
5th best route	4.98	1.05	9.37	6.68	24.27	19.09
6th best route	6.38	1.55	10.90	7.77	25.68	20.55
7th best route	7.55	1.50	13.35	7.60	34.20	24.39

3.1.2 Route Images Generation on World Maps

Around two hundred seventy routes were generated on a 100×100 grid (Figure 15) in Wolfram Mathematica by calculating the shortest distance between a fixed start and endpoint using Dijkstra's shortest path algorithm (Dijkstra, 1959). The shortest distance changes for each route because some portion of the grid was randomly deleted from the

shortest-path calculation in every iteration of route generation. Seven routes were randomly picked from all 270 routes for each trial. The only constraint for selecting seven routes was that they are spread out on the grid and have little or no overlap to avoid confusion by participants in distinguishing them. The x and y coordinates for each vertex of the routes were converted to distance in miles for different start and end positions on a world map defined using their longitude and latitude values. The seven routes were drawn on the world map using the position for each vertex (Figure 16). The latitude and longitude values for each start and end destination for all 40 trials were randomly extracted using the Classic Searoutes website.

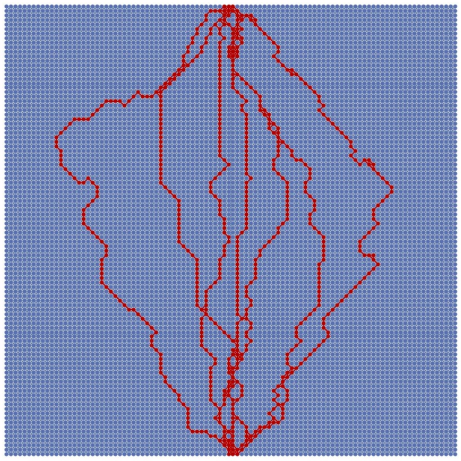

Figure 15. Seven routes generated on a 100×100 grid in Mathematica

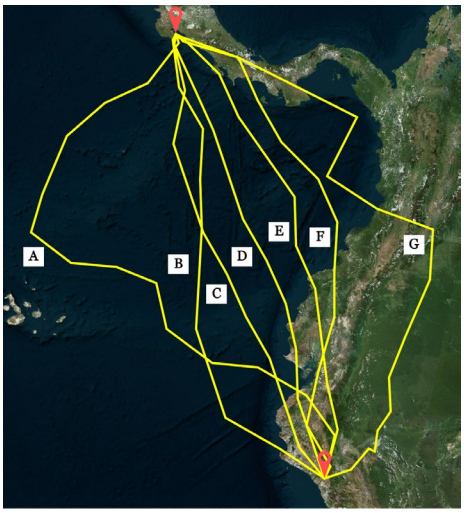

Figure 16. Routes generated in the grid are superimposed on a world map and labeled from A-G

3.2 Experiment 1: Investigating the Impact of Model Blindness on Decision-making

The goal of the first study in this dissertation was to show how misspecifications in MDSS' model impose model blindness on users and how that impact decision-makers' performance, decision-making process, trust, and confidence in the system. Another important goal was to show how models can bias human-decision making and enhance the impact of model blindness by presenting a few salient alternatives to the users. The experiment was designed to help understand how model-limited and strategy-limited performance can manifest in this study's route-planning task involving an unfriendly environment. Both model misspecification and the presence of recommended alternatives were manipulated systematically to investigate this issue of model blindness.

3.2.1 Design

The experiment was designed using PsychoPy software (Peirce et al., 2019; Peirce et al., 2022) for participants to perform a route-planning task using the route-recommender system described in the previous section. The experiment was a 2 (Route Recommender System) x 3 (Model Misspecification) between-subjects design, as shown in Figure 17. The recommender system manipulation had two levels: (1) three MDSS-preferred recommended routes absent (Figure 13), and (2) three MDSS-preferred recommended routes present (Figure 18). The model misspecification manipulation had three levels: (1) accurate model (no misspecification), (2) soft misspecified model, and (3) hard misspecified model. The soft misspecified model overweighted the attribute "Time Efficiency." The hard misspecified model overweighted the attribute "Time Efficiency" and "Obstacle Avoidance." The dependent measures included performance (response rank and utility loss), trust score, calibration (Brier score), reliance on preferred routes, and participants' route choices (used to extract decision strategy). All the conditions with preferred routes absent served as control conditions for the corresponding experimental conditions with preferred routes present. The experiment required three separate control groups for the corresponding experimental groups because the set of seven routes and corresponding decision attribute values were different based on the level of misspecification in the MDSS, which determined the quality of the whole alternative set and recommended set (Table 3). So, comparing the hard misspecified experimental condition to an accurate control condition was inappropriate. Because even though participants in the hard misspecified condition received additional decision support from the recommended set (top 3 routes), the quality of all seven alternatives presented was

much better for the accurate control condition compared to the hard experimental condition on average. Also, the recommender was much more accurate in no misspecified conditions.

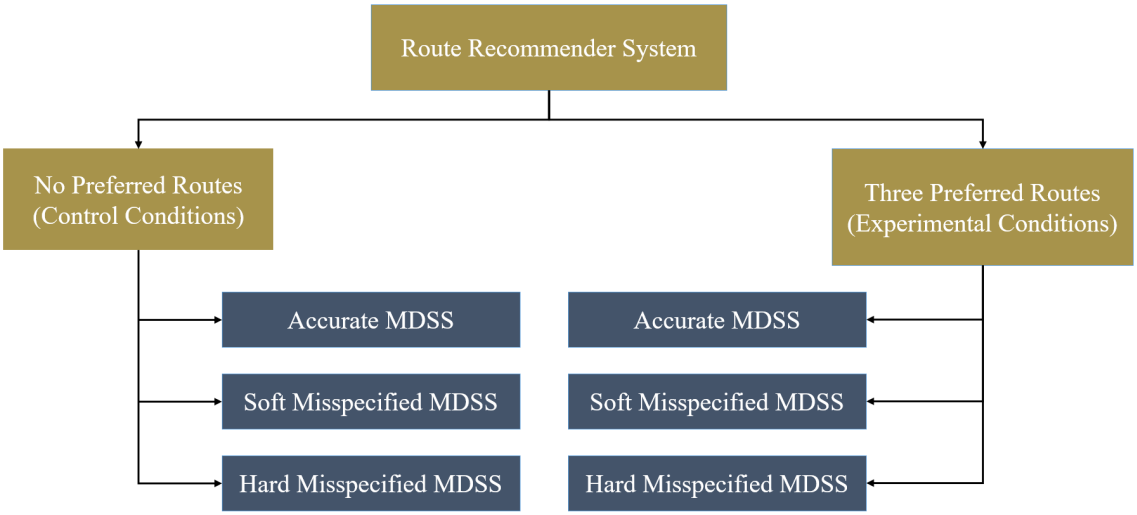

Figure 17. Design for Experiment 1

In order to prevent any order effects in the experiment, I randomized every aspect of the experiment. Participants were randomly assigned to the six experimental conditions. The order in which the six attributes were presented was randomized across participants, but the order was constant for the same participant to avoid confusion in attribute label reading. The route labels (A-G) in the 40-trial images were also randomized using a balanced Latin square (Online Calculator). The route labels stayed the same for the same trial image for all six conditions. The preferred routes (top 3 identified by each MDSS) were colored differently (solid yellow vs. dashed red in Figure 18).

Figure 18. Route Recommender System interface for experimental groups in Experiment 1 and 2 (with system preferred route set represented as solid yellow lines and additional recommended routes represented as dashed red lines)

3.2.2 Participants

193 participants were recruited to participate in Experiment 1. Participants for this study were recruited from the Georgia Tech Sona experiment management system and were compensated with 1.5-course credit. Any student at Georgia Tech with normal or corrected-to-normal vision was eligible to participate in Experiment 1. I planned to recruit at least 180 participants for Experiment 1, i.e., at least 30 for each condition. The proposed sample size was obtained by performing an a-priori power analysis in GPower software. Given this proposed sample size and assuming an alpha of 0.05, I anticipated being able to detect a medium effect size of approximately cohen's f = 0.31 (cohen's d = 0.62) with 80% power. The data collection for this experiment began on May 17th, 2022. Before beginning

data collection, the study was pre-registered, and pre-registration is available on OSF (https://osf.io/p5dkb).

3.2.2.1 Exclusion Criteria

The data for 180 participants were included in the analysis. The remaining participants were excluded for the following reasons: pilot participants (5), participants with more than 10% (4 trials) of trial response data missing (4), participants who needed a long break during the experiment (2), and participants who accidentally exited code (2). Block randomization was used to achieve an equal minimum number of participants in each condition. Data collection was paused after accommodating exclusion criteria, when I reached the target number of participants in each condition.

3.2.2.2 Participant Demographics from Pre-Study Questionnaire

I only have demographics questionnaire responses for 158 participants (88% participants) as the pre-study questionnaire (See APPENDIX B) was added to the study after beginning data collection for this study. The mean age for 156 participants (2 didn't report) was 19.23 years (Range: 18-26 years). The gender distribution for participants is presented in Table 5, and their major distribution at Georgia Tech is presented in Figure 18. Computer Science was the most represented major among participants. The majority of participants (78%) had past experience playing video games (No experience- 10%, Missing Data- 12%). The majority of participants (82%) also had past experience with AI-based recommender systems (e.g., Netflix movie recommendations, Amazon product recommendations, Health and Fitness apps) in their day-to-day life (No experience- 4%, Missing data- 14%). A small number of participants (12%) had also taken a

course/worked with recommender systems in the past (No course- 73%, Missing data- 14%).

Table 5. Gender distribution of participants in Experiment 1

Gender	Percentage of Participants
Female	31%
Male	54%
Non-binary/Non-conforming / Other	2%
Didn't report	1%
Missing Data	12%

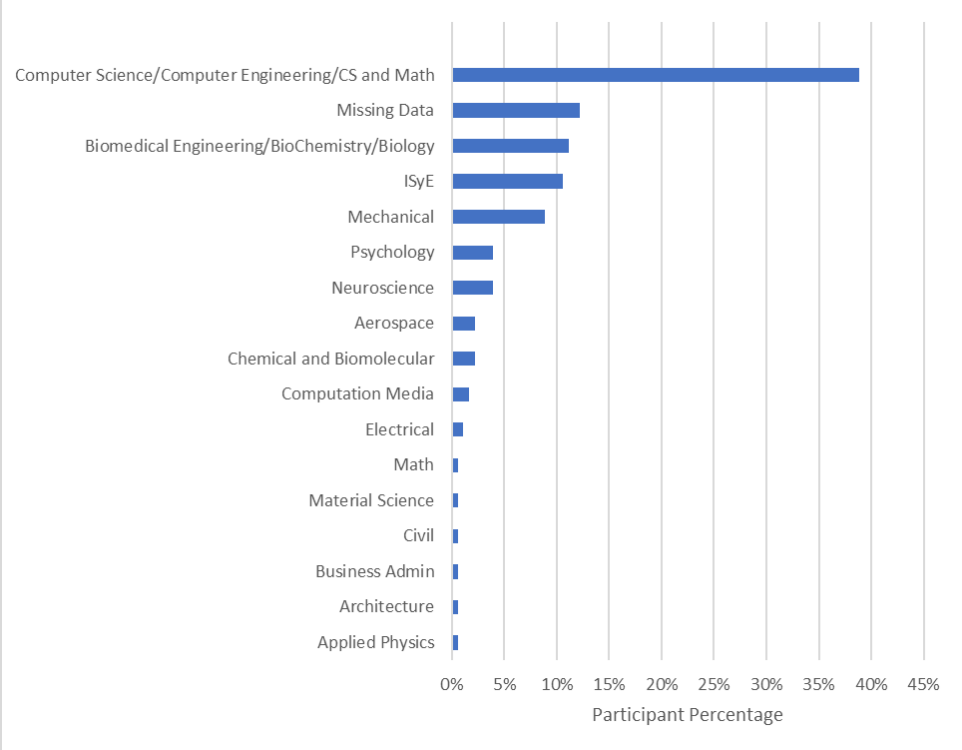

Figure 19. Major distribution of participants in Experiment 1

3.2.3 Procedure

Participants were seated at a desk with a computer in Decision Processes Lab at the School of Psychology at Georgia Tech to complete all the tasks involved in the experiment. The entire experiment, including the consent form at the beginning and the debriefing at the end of the experiment, was administered using PsychoPy software installed on participant computers.

As shown in Figure 20, the experiment protocol started when participants provided informed consent at the start of the experiment and background demographics information (See APPENDIX B for the pre-study demographics questionnaire). Following that, they received instructions about the route-recommender system and what decision task they had to perform (See APPENDIX A for detailed participant instructions). Participants were instructed that they would be provided with an AI-based route recommender system that would present seven routes and corresponding attributes where a higher attribute value would mean better value (or utility). The participants would click on button A-G, shown on the right of route images in Figure 13 (control group) or Figure 18 (experimental group) to look at attributes corresponding to each route (see Experiment Video).

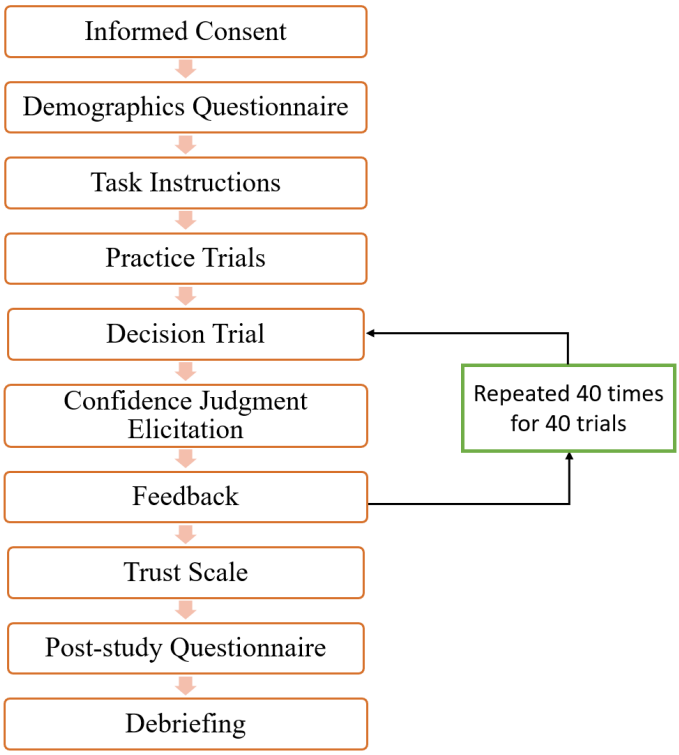

Figure 20. Procedure steps for Experiment 1

The system was presented as AI-based to the participants due to the familiarity of the general population with this term in everyday life compared to an MDSS. As remote operators, they had to perform a route planning task (sea + ground shipping) to deliver critical care shipments from one location to another in urgent need using the provided AI-based route recommender system. At the start of the experiment, the initial instructions explained in detail what the route recommender system was and what each decision attribute meant. Participants were instructed to use all the presented information and that each piece of information was equally important for an optimal decision. By instructing participants to use all information equally, I wanted to prime participants to use an equal weights decision strategy as the true model underlying the route-recommender system was an equal weights model. The participants received four dummy learning trials at the start

of the experiment to learn how the route-recommender system interface worked. After that, participants were presented with 40 trials on which they had to choose the most optimal route out of the seven. After each trial, participants were asked to rate their confidence in their decision (Figure 21). Following that, they received feedback (Figure 22) regarding where their chosen route ranked out of the seven routes on which they made their decision choice. The feedback was aimed at helping participants learn what attribute weights the true model uses to pick the most preferred alternative. At the end of the experiment (after 40 trials), participants were administered a trust scale measuring their trust in an AI-based recommender system. After the trust scale, participants responded to the post-study questionnaire determining their experience with the study and their past experience with recommender systems in general (See APPENDIX C for the post-study questionnaire). Finally, in the end, participants were provided a debrief summary of what this experiment was trying to study.

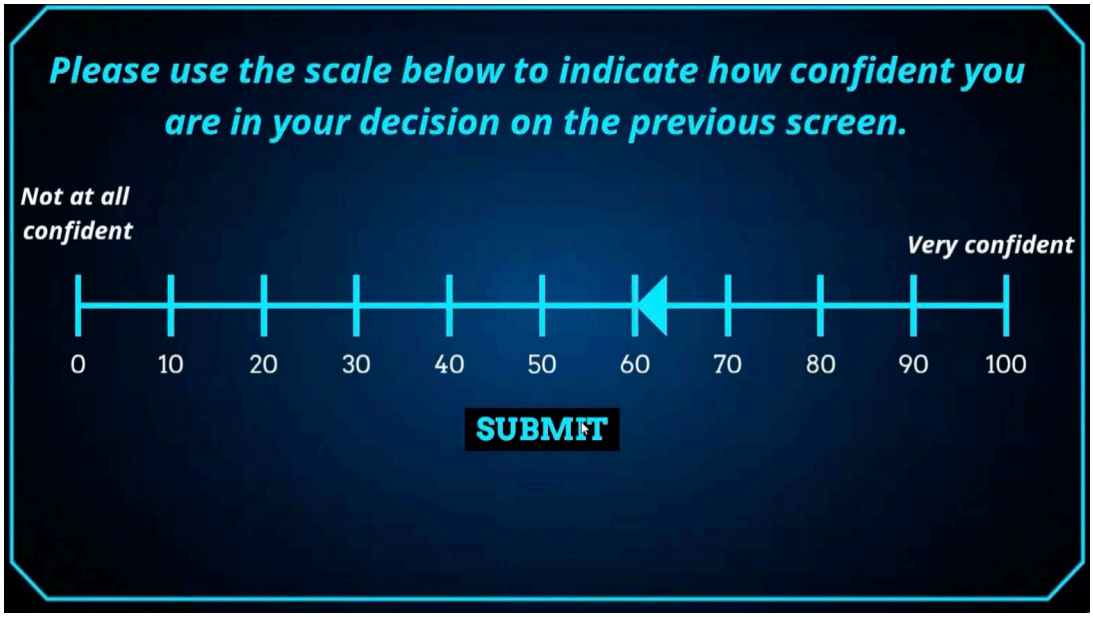

Figure 21. Confidence judgment elicitation

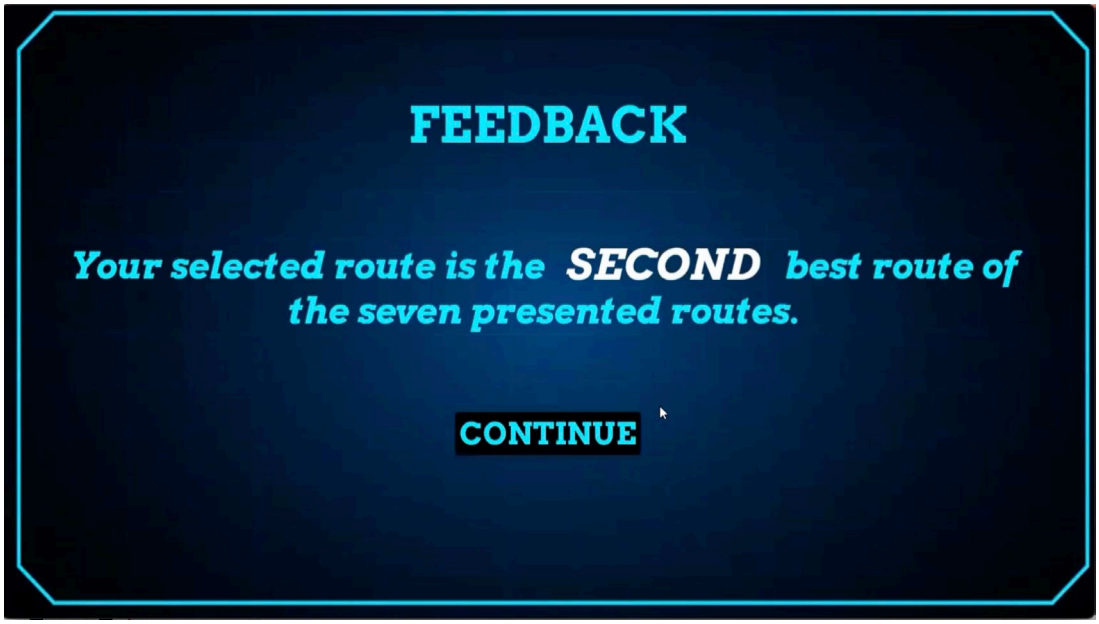

Figure 22. Feedback presented to participants after each trial

3.2.4 Validation of Task Ecology through Simulation

Before beginning participant data collection, I investigated whether the ranked feedback presented in this experiment had enough statistical properties for participants to learn task ecology (attribute weights and inter-attribute correlation) over the course of the experiment (a total of 40 trials) through simulations. I performed ordinal logistic regression with ranked feedback as the criterion and all six decision attributes as predictors using R software for statistical analysis.

A simulated decision-maker randomly picked one route alternative out of 7 on all 40 trials, which yielded one random response sample. Then ordinal regression was run

using the feedback rank and corresponding attribute weights for the random sample generated by a simulated decision-maker. Due to high multicollinearity among decision attributes, ordinal regression failed to converge for many random samples. To check the likelihood of the regression model not converging, I ran 100 random samples for each model, including hard, soft, and accurate. The hard model failed to converge 23 times, the soft model failed to converge 8 times, and the accurate model always converged.

Table 6. Ordinal regression weights for 20 random samples of decision choices (route-choices of 20 random simulated participants on 40 trials)

MDSS Model	Average beta weights (Number of samples with $p<0.05$)					
	β_1	β_2	β_3	β_4	β_5	β_6
Accurate Model	-0.038 (16)	-0.035 (16)	-0.039 (15)	-0.036 (14)	-0.041 (17)	-0.033 (15)
Soft Misspecified Model	-0.0004 (0)	-0.063 (20)	-0.063 (19)	-0.059 (20)	-0.055 (19)	-0.06 (20)
Hard Misspecified Model	-0.01 (1)	-0.07 (20)	-0.016 (3)	-0.06 (20)	-0.067 (20)	-0.07 (20)

For each model, I randomly picked 20 random samples that converged on ordinal regression out of 100 samples. That can be assumed to be a response from 20 simulated decision-makers acting randomly during the experiment in all three control conditions. Table 6 shows the average beta weights for 20 random samples with 40 decision choices each. The table also shows how many samples, out of 20, the beta weight was significant

($p<0.05$). For the accurate model, all the beta weights (standardized regression coefficients) were almost equal in magnitude, indicating that each attribute value must be weighted equally to achieve lower ranked feedback from those values, thus indicating the use of equal weights true model for optimal performance. Also, most attributes were significant on 14-17 samples out of 20. However, not all of them were statistically insignificant in the same sample. The number indicates that for some random samples, one or two attributes were statistically insignificant. This might have happened because of the presence of multiple highly correlated predictors that the decision-maker might not need information from all sources to succeed in a task, i.e., the value on one attribute can compensate for the value on another.

For the soft misspecified model, the beta weight for the first attribute (β_1) was not statistically significant in any of the 20 samples, and its magnitude was also much lower compared to the other five attributes indicating that the participants would have to underweight or completely ignore this attribute to select a lower ranked (best route) route alternative. They would also have to give equal value to the other five attributes to perform well. This again validated the weights that participants would have to put into selecting the best alternative as the soft misspecified MDSS overweighted (double) the first attribute while generating and presenting the top 7 routes to the participants. The participants would have to compensate for this overweighting via MDSS by underweighting that attribute and weighing others equally.

For the hard misspecified model, the beta weights for the first attribute (β_1) and the third attribute (β_3) were not statistically significant for most of the 20 random samples, and their magnitude of beta values was also much lower compared to the other four attributes

indicating that the participant would have to underweight or completely ignore these attributes to select a lower ranked alternative (best route). They would also have to give equal value to the other four attributes to perform well in the route planning task. This again validated the weights that participants would have to put into selecting the best alternative as the hard misspecified MDSS overweighted (double) the first and the third attributes while generating and presenting the top 7 routes to the participants. Participants would have to compensate for this overweighting via MDSS by underweighting those two attributes and weighing others equally.

These results helped validate that the ranked feedback had enough statistical information to learn and understand how to weigh attributes appropriately during the experiment. Hence, these results served as motivation to start data collection with ranked feedback in the study.

3.2.5 Dependent Measures in the Study

The following sub-sections discuss different dependent measures used in this study and the reasoning for using them as dependent variables.

3.2.5.1 Performance Measures

Response rank was measured as the selected route's ordinal rank out of the seven presented routes by the MDSS on each trial. It varies from 1 to 7. It mimics the ranked feedback that participants received after each trial.

Outcome was measured as a binary variable indicating whether participants picked the top-ranked route or not. In each trial, participants received a score of 1 when they picked the first best route and received a score of 0 when they picked any other route.

Global utility loss was based on the utility of the true best alternative produced by the true equal weights model (using true weights and attribute values without any noise). The global utility loss was measured using (see Equation 3) using the utility of the true best alternative out of 150 routes used by the true model to develop the MDSS.

$$Global\ Utility\ Loss = \frac{Utility_{Best\ Alternative(out\ of\ 150\ routes)} - Utility_{Selected\ Alternative}}{Utility_{Best\ Alternative(out\ of\ 150\ routes)}} \quad (3)$$

Local utility loss was based on the true utility of the best alternative presented by the MDSS to participants in each trial. It is derived by using the global utility loss values available for each route using Equation 4.

$$Local\ Utility\ Loss = \frac{Global\ Utility\ Loss_{\substack{Selected\\Alternative}} - Global\ Utility\ Loss_{\substack{Local\ Best\\Alternative\\(out\ of\ 7\\presented\ routes)}}}{Global\ Utility\ Loss_{\substack{Local\ Worst\\Alternative\\(out\ of\ 7\\presented\ routes)}} - Global\ Utility\ Loss_{\substack{Local\ Best\\Alternative\\(out\ of\ 7\\presented\ routes)}}} \quad (4)$$

Response rank, outcome, and local utility loss are MDSS-centric performance measures in which participants' performance was evaluated based on the alternatives they received. These measures solely assessed performance based on the information that participants received from MDSS during the experiment. With response rank being similar to the ranked feedback that participants received, the outcome being equivalent to participants' absolute success on each trial, and local utility loss evaluated participant

performance based on the absolute true accuracy of the tool. On the other hand, global utility loss was based upon the true best alternative produced by the true equal weights model defining the true state of the world for the MDSS used in the study.

3.2.5.2 Confidence Judgement and Calibration

Participants rated their confidence in their route choice after each trial decision on a 0 (not at all confident) to 100 (completely confident) continuous scale. Per trial confidence judgments of participants was used to measure the calibration of participants' confidence to the actual outcome of their decision using Brier scores (BS) (see Equation 5) as analyzed by Parmar and Thomas (2020). Brier score is the most common metric for the analysis of confidence judgment (Brier, 1950; Murphy & Winkler, 1977; Yates, 1990). The Brier score (BS) is a proper scoring rule that provides a measure of the accuracy of confidence judgments. Hence, in this case, it helped measure how much participants' confidence judgments were calibrated with the actual MDSS accuracy.

$$BS = \frac{1}{N}\sum_{T=1}^{N}(c_t - o_t)^2 \tag{5}$$

The Brier score is described by Equation 5, where N is the total number of probability or confidence assessments, c_t is the t^{th} confidence judgment, and o_t is the outcome index for the t^{th} confidence judgment. If the event occurs (if a participant selects the correct route out of seven), then $o_t = 1$, and if the event does not occur (if a participant selects the incorrect route out of seven), then $o_t = 0$. The outcome index is equivalent to the

outcome measure described previously as one of the performance measures. Thus, the Brier score is the average squared deviation between the confidence of the decision-maker and the outcome index (Brier, 1950; Murphy & Winkler, 1977; Yates, 1990). The lower the Brier score for a set of predictions, the better the predictions are calibrated (i.e., less error in predictions).

3.2.5.3 Reliance and Trust in MDSS

Participants' reliance on MDSS was derived as a binary measure only for experimental conditions. If participants chose a route alternative from the set of three preferred routes (solid yellow lines in Figure 18), their reliance was scored as 1 on that trial. If participants chose a route alternative from the set of four non-preferred routes (dashed red lines in Figure 18), their reliance was scored as a 0 on that trial.

Reliance on any automated system depends on users' trust in the system (Lee & See, 2004). Therefore, the participants' trust in the MDSS was also measured at the end of the experiment. The participants were administered a 4-point Likert-based trust scale (Table 7) developed by Ashoori and Weisz (2019) to measure trust in the AI-based DSS. The authors developed this scale by re-wording items and pooling from multiple scales (including (Jian et al., 2000)) already available in the trust in the automation literature that evaluates different aspects of user trust. Ashoori and Weisz (2019) developed a separate scale, as most existing trust scales focus on autonomous systems rather than AI-based or model-based systems that provide decision support to the participants.

Table 7. Items of trust scale by Ashoori and Weisz (2019) grouped by the facet of trust the items focus on (*reverse scored items)

Trust Facet	Trust Scale Items
Trustworthiness	This decision-making process is trustworthy
	I would change one or more aspects of this decision-making process to make it trustworthy*
	This decision-making process will produce a fair outcome for the person affected by the decision
	The decision-maker needs more information about how the AI model was trained and tested in order to trust the process*
Reliability	This decision-making process would always make the same recommendation under the same conditions
	The outcome of this decision will be consistent with other decisions made for other people
Technical Competence	The use of an AI model is appropriate in this scenario
	This decision will be made based on reliable information
	I trust that the technical implementation of the AI model is correct
Understandability	It is easy to understand what this decision-making process does
	I understand how this decision-making process works
Personal attachment	I am confident in this decision-making process. I feel that it works well
	I am wary of this decision-making process *
	I like this decision-making process

3.2.6 Analysis and Results

Linear mixed-effects models were used to analyze all continuous repeated measure DVs (response rank, utility loss, Brier score, and confidence judgments), and generalized linear mixed-effects models were used to analyze repeated measure binary DVs (reliance and outcome) for the complete experiment design with participants as the random effects grouping factor. A two-way ANOVA with post hoc comparisons was used for trust scores, which were measured once per participant.

For all my DVs, I expected to find significant main effects of both independent variables: recommender [conditions: yes (with recommender preferred 3 routes vs. no (without recommender preferred 3 routes)] and model misspecification (conditions: accurate, soft, and hard). I also expected to see a significant interaction between model misspecification and recommender. However, little to no evidence was found for the main effect of model misspecification and an interaction between recommender and model misspecification for most of my DVs. For most DVs, there is only evidence for the main effect of the recommender. The following sub-sections discuss the results organized by DVs.

3.2.6.1 Performance Measures

Response rank

Figure 23 shows the mean rank of participant route choices by each condition in the experiment. The response rank is based on the true model (out of 7 presented routes) and mimics feedback received by participants on each trial. A lower rank means better

performance on trial as the experiment task was to find the 1st ranked route on each trial. A repeated measures ANOVA using a linear mixed-effects model was used to test the effect of recommender availability and model misspecification on participant decision choices using response rank as a continuous variable. The trial ID variable, which represents the 40 different trials (scenarios) that participants received in each condition, was also included as a predictor in the model to test if the performance varies significantly across different trials. It is important to remember that participants in different misspecification conditions received different attribute values for the same trial ID due to misspecification, however the true model is same for each trial ID across conditions.

The ANOVA summary table is shown in Table 8. Participants response ranks differed significantly between recommender conditions (χ^2 (1) = 6.23, p = .013). The response rank for participants with recommender was lower (better performance) compared to participants without recommender (Mean Difference= 0.172). The main effect of trial ID (χ^2 (39) = 859.78, p < 0.001) was also significant, indicating a difference in their difficulty level affecting performance. The interaction between trial ID and model misspecification (χ^2 (78) = 670.99, p < 0.001) was also significant, indicating that performance on different trials varied significantly across conditions. Some trials were more difficult than others at some misspecification levels (Figure 24).

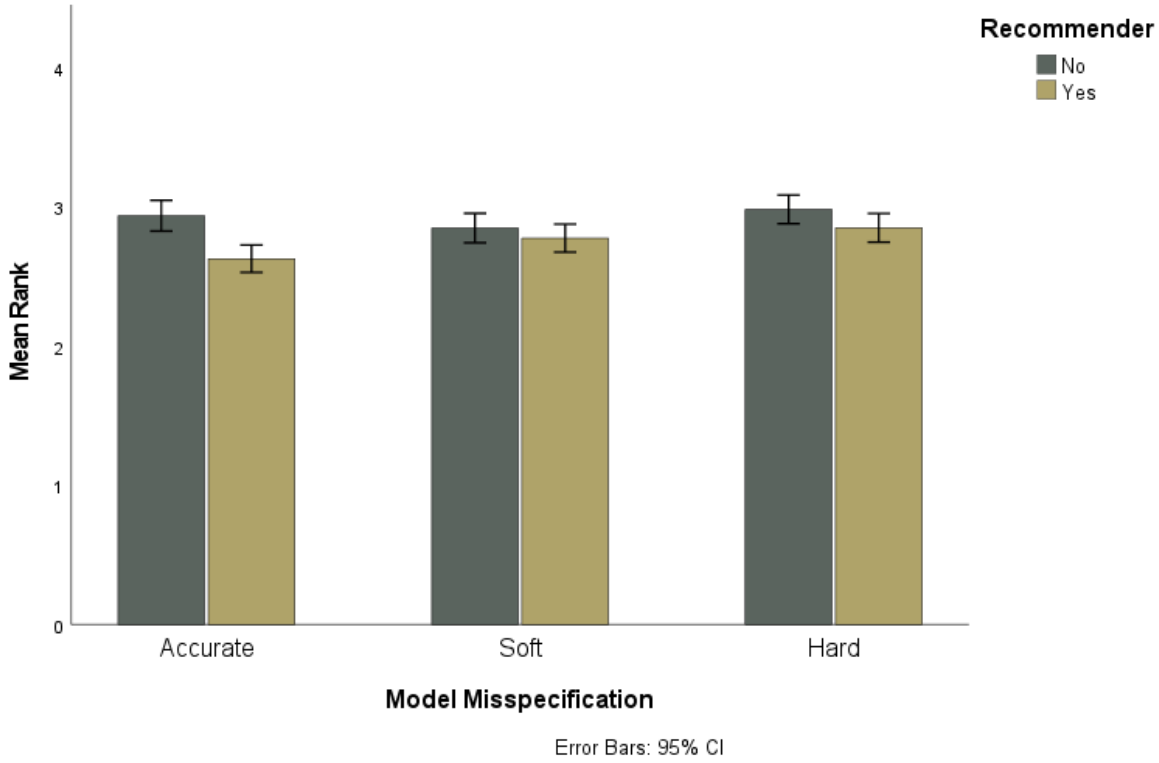

Figure 23. Mean response rank (out of 7) of selected route by model misspecification and recommender conditions

Table 8. Linear mixed models: Likelihood ratio tests for response rank

Effect	df	ChiSq	p
Model Misspecification	2	2.839	0.242
Recommender	1	6.226	0.013
Trial ID	39	859.795	< .001
Model Misspecification × Recommender	2	2.153	0.341
Model Misspecification × Trial ID	78	670.988	< .001
Recommender × Trial ID	39	38.847	0.477
Model Misspecification × Recommender × Trial ID	78	85.922	0.252

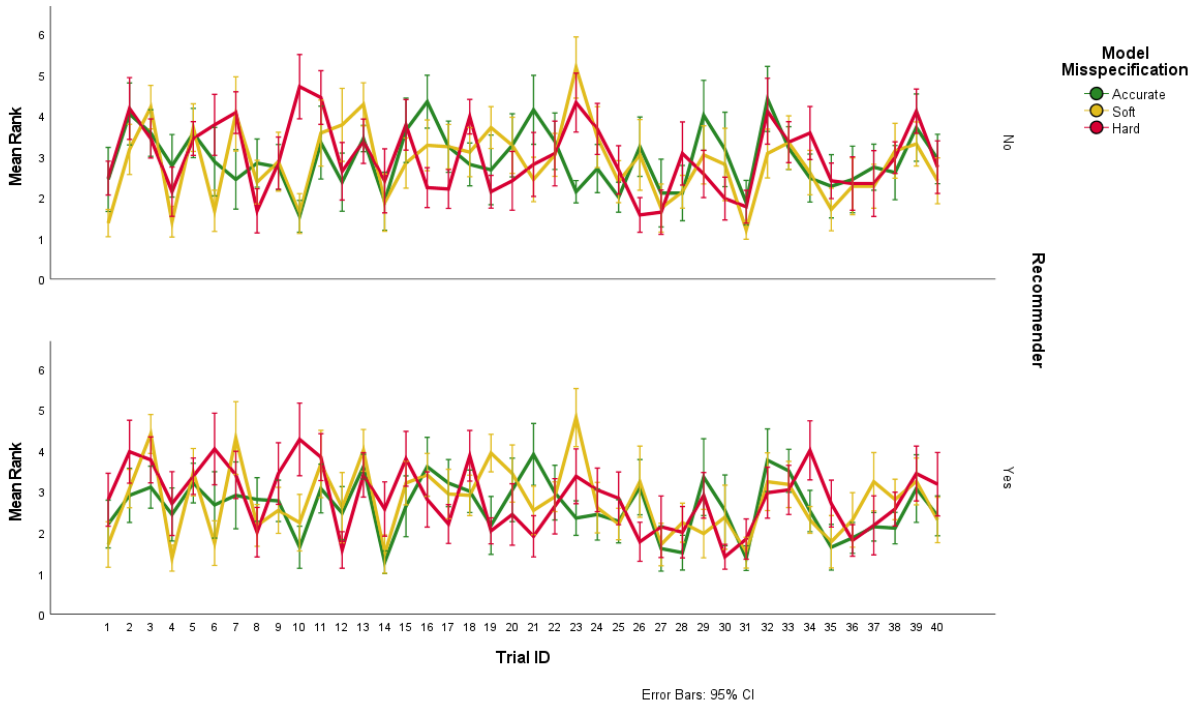

Figure 24. Mean response rank (out of 7) of the selected route by model misspecification and recommender conditions for all trial IDs

There was no significant main effect of model misspecification or interaction between model misspecification and recommender. I predicted that participants in accurate and hard recommender conditions would perform better than participants in their corresponding no recommender (control) conditions. This was because of the additional support provided by the recommender system that displayed highly accurate routes in the recommended three routes. The prediction for hard condition was exploratory. I predicted the trend because participants received a highly misspecified MDSS in the hard recommender condition, which might make it easier to detect that the recommender suggests less valuable alternatives, and participants can deviate from the recommender and use their own judgment or recommended EQ weights policy to make the route-decision.

The evidence for hard and accurate conditions can be seen from significant main effect of recommender. I also predicted that participants in the soft recommender condition would perform worse than the corresponding soft control condition. However, no evidence was found for this prediction. This prediction was exploratory and counterintuitive. I expected this trend because participants might find it difficult to overcome slightly misspecified support by a recommender. Participants might take longer (more trials) to overcome the soft misspecified recommender until they realize that the recommender system is not very accurate.

For all control conditions, I expected performance to worsen with increasing misspecification from accurate to soft to hard. This was because the quality of the alternative set degraded with the increase in the MDSS level of misspecification. For all recommender conditions (experimental), I expected that the order of performance degradation would be first accurate, then hard, then soft condition. This is again because of difficulty overcoming a slightly misspecified system (soft) that has a higher probability of being correct sometimes compared to a highly misspecified system (hard), which has a high probability of being wrong most of the time.

Outcome

The outcome measure was coded as a binary variable with a score of 1 on trials when the best route was selected and 0 on trials with any other ranked route selection. This DV was implemented to test if a more conservative scoring approach reveals performance differences across conditions. All the predictions for outcome DV were the same as the response rank DV. The proportion of trials with outcome equal to 1 for all conditions is

shown in Figure 25. A repeated measures ANOVA using a generalized linear mixed effects model with binomial family and logit link function was used to test the effect of model misspecification, recommender availability, and trial block on the outcome. The trial block was used instead of trial ID as a predictor in generalized linear mixed effects model as the large number of levels of trial ID cannot be estimated by the model for available data. The trial block is ordinal variable with 4 levels with each level comprising 10 trials in order of presentation. No significant effects were found (Table 9). Hence, no evidence was found for the predicted main effect of model misspecification, recommender availability, and an interaction between the two. No significant difference in participants' ability to select the best outcome changed with the level of misspecification and/or recommender availability.

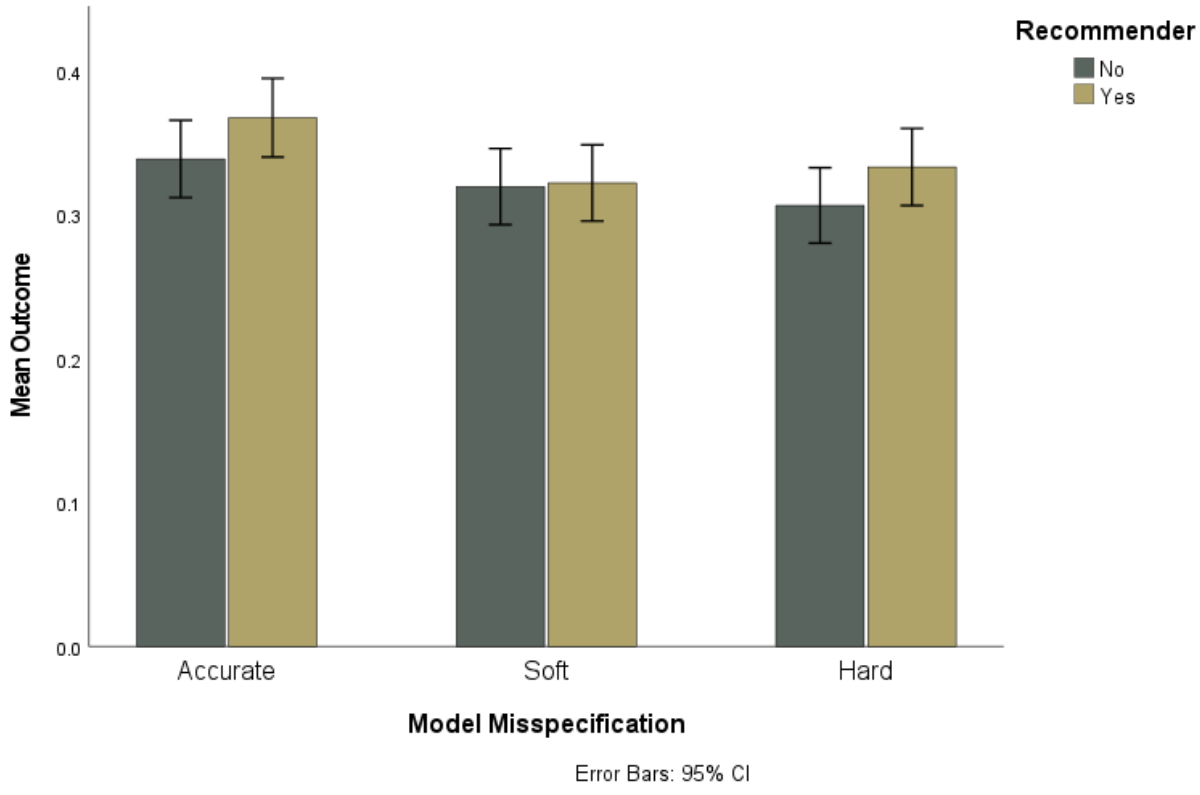

Figure 25. Mean outcome (proportion of trials with the best route selected) by model misspecification and recommender conditions

Table 9. Generalized linear mixed models: Likelihood ratio tests for outcome

Effect	df	ChiSq	p
Model Misspecification	2	3.527	0.171
Recommender	1	1.500	0.221
Trial Block	3	0.647	0.886
Model Misspecification × Recommender	2	0.497	0.780
Model Misspecification × Trial Block	6	5.690	0.459
Recommender × Trial Block	3	3.002	0.391
Model Misspecification × Recommender × Trial Block	6	6.539	0.366

Local utility loss

The local utility loss (Figure 26) is again a performance measure with the same predictions as response rank and outcome. The lower utility loss is considered better. A repeated measures ANOVA using a linear mixed-effects model was used to test the effect of recommender availability and model misspecification on participant decision choices using local utility loss as a continuous variable. The trial ID was also included as a predictor in the model to test if the local utility loss varies significantly across different trials. The ANOVA summary table is shown in Table 10. The significant effects are similar to the response rank effects presented before. Participants' local utility loss for their selected routes differed significantly between recommender conditions (χ^2 (1) = 5.62, p = .018). The local utility loss for route choices of participants with recommender was lower (better performance) compared to participants without recommender (Mean Difference= 0.031). The main effect of trial ID (χ^2 (39) = 909.92, $p < 0.001$) was also significant, indicating a difference in trial difficulty level affecting performance. The interaction between trial ID and model misspecification (χ^2 (78) = 988.37, $p < 0.001$) was also significant, indicating that performance on different trials varied significantly across conditions. Some trials were more difficult than others at some misspecification levels. The local utility loss and

response rank measures were normalized (z-scores) by Trial ID to account for item difficulty, and the same ANOVA models were rerun. However, the results stayed the same, except the main effect of Trial ID was non-significant, as expected due to normalization.

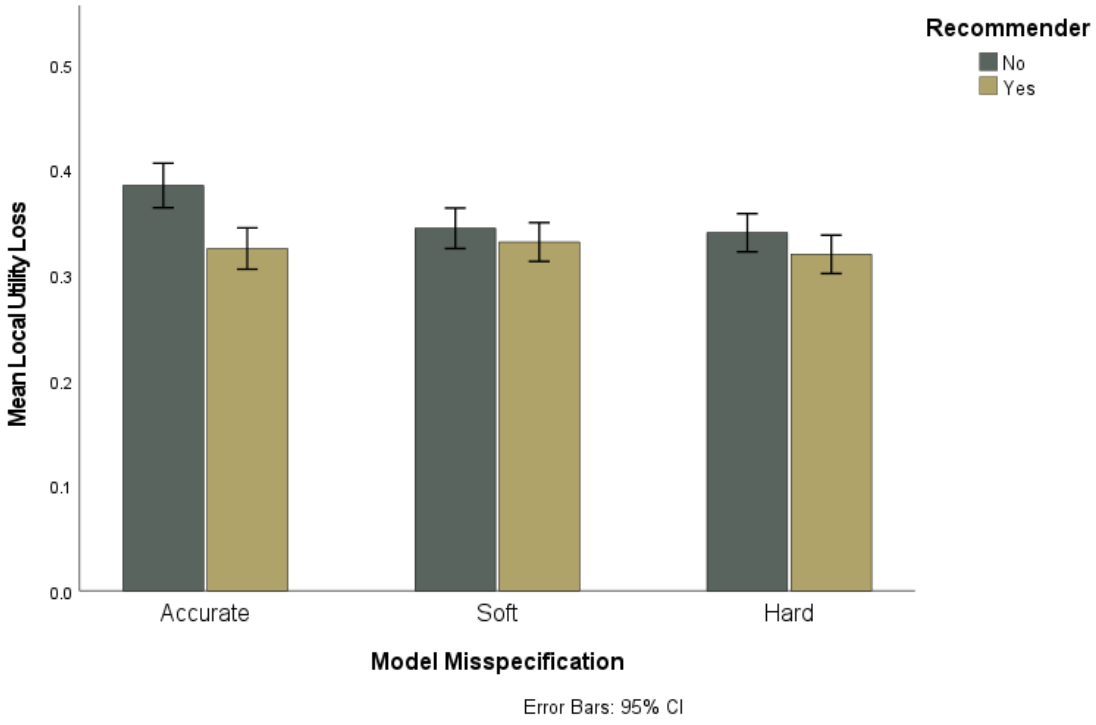

Figure 26. Mean local utility loss of selected route by model misspecification and recommender conditions

Table 10. Linear mixed models: Likelihood ratio tests for local utility loss of selected route

Effect	df	ChiSq	p
Model Misspecification	2	2.597	0.273
Recommender	1	5.627	0.018
Trial ID	39	909.922	< .001
Model Misspecification × Recommender	2	2.592	0.274
Model Misspecification × Trial ID	78	988.372	< .001
Recommender × Trial ID	39	39.129	0.464
Model Misspecification × Recommender × Trial ID	78	72.591	0.652

Global utility loss

The global utility loss (Figure 27) is again a performance measure like local utility loss, with a lower value being considered better. However, global utility loss evaluates participants' route choice in comparison to the true global best route generated by the true equal weights model. It compares route choices to the true world for MDSS instead of just seven presented alternatives. As it represents the true world, I expected it to follow the order of misspecification (main effect) for both levels of the recommender. The global utility loss should increase as the quality of the alternative set gets worse with increasing misspecification. However, global utility loss measure performance in comparison to true global best alternative and hence, should be unrelated to local task-specific measures like trust, confidence, and reliance. A repeated measures ANOVA using a linear mixed-effects model was used to test the effect of recommender and model misspecification on participant decision choices using global utility loss as a continuous variable. The trial ID was also included as a predictor in the model to test if the global utility loss varies significantly across different trials. The ANOVA summary table is shown in Table 11. Participants' global utility loss for their selected routes differed significantly between model misspecification conditions (χ^2 (2) = 188.12, $p < .001$). The post-hoc contrasts show significant differences in global utility loss between all three groups in order of misspecification:1) accurate condition has lower global utility loss compared to soft (MD=0.014; z = 5.16; p<0.001), 2) accurate condition has lower global utility loss compared to soft (MD=0.049; z = 17.71; p<0.001), and 3) soft condition has lower global utility loss compared to hard (MD=0.034; z = 12.56; p<0.001).

The main effect of trial ID (χ^2 (39) = 1512.28, $p < 0.001$) was also significant, indicating a difference in their difficulty level affecting performance. The interaction between trial ID and model misspecification (χ^2 (78) = 931.76, $p < 0.001$) was also significant, indicating that performance on different trials varied significantly across misspecification levels. Some trials were more difficult than others at some misspecification levels. The interaction between trial ID and recommender (χ^2 (39) = 55.26, $p = 0.044$) was also significant, indicating that performance on different trials varied significantly between both recommender conditions.

Table 11. Linear mixed models: Likelihood ratio tests for global utility loss of selected route

Effect	df	ChiSq	p
Model Misspecification	2	188.125	< .001
Recommender	1	2.780	0.095
Trial ID	39	1512.280	< .001
Model Misspecification × Recommender	2	0.685	0.710
Model Misspecification × Trial ID	78	931.764	< .001
Recommender × Trial ID	39	55.261	0.044
Model Misspecification × Recommender × Trial ID	78	92.814	0.121

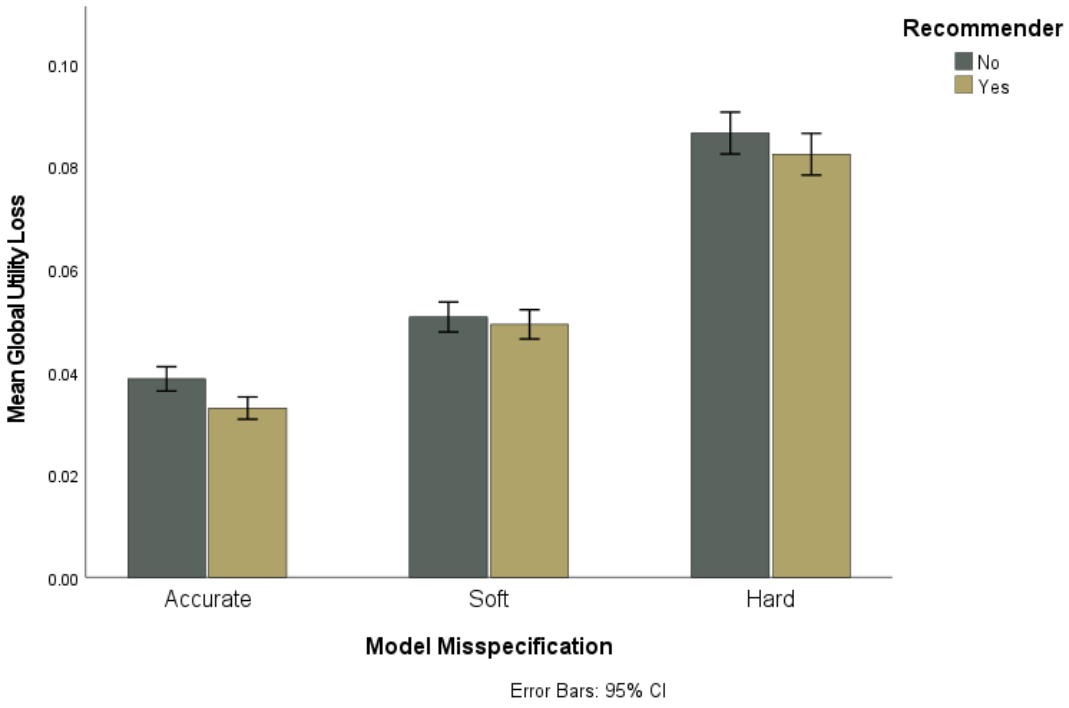

Figure 27. Mean global utility loss of selected route by model misspecification and recommender conditions

3.2.6.2 Confidence Judgement and Brier Scores

Although I expected the Brier scores of the participants' confidence judgments to follow a similar pattern as performance, the Brier Scores were expected to be less sensitive to misspecifications in the MDSS than performance because the confidence elicitation was on a per-trial basis before participants received performance feedback. Moreover, performance was more dependent on the MDSS's absolute accuracy, while confidence was based on the more subjective assessment of the decision makers' choice.

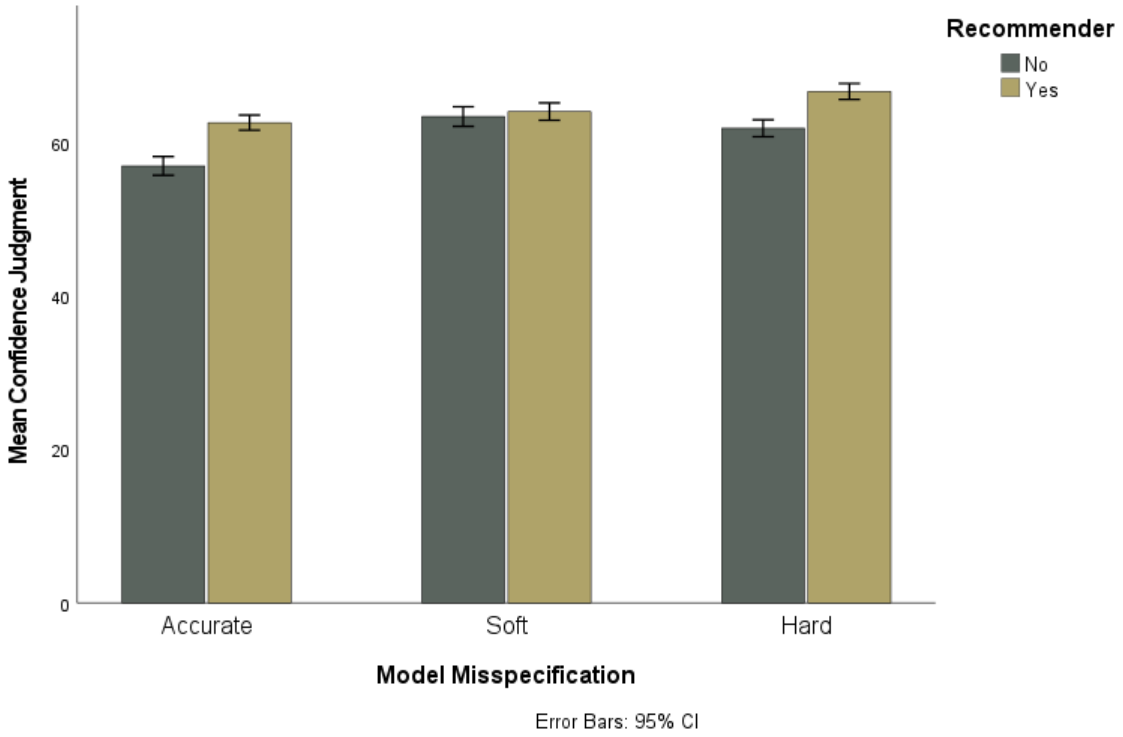

Figure 28. Mean confidence judgment score by model misspecification and recommender conditions

Two repeated measures ANOVA using a linear mixed-effects model were conducted to test the effect of recommender availability and model misspecification on participant confidence judgments (Figure 28) and Brier scores (Figure 29). The trial ID was also included as a predictor in both models to test whether confidence judgments or Brier scores change significantly across different trials. The ANOVA summary tables are shown in Table 12 and Table 13.

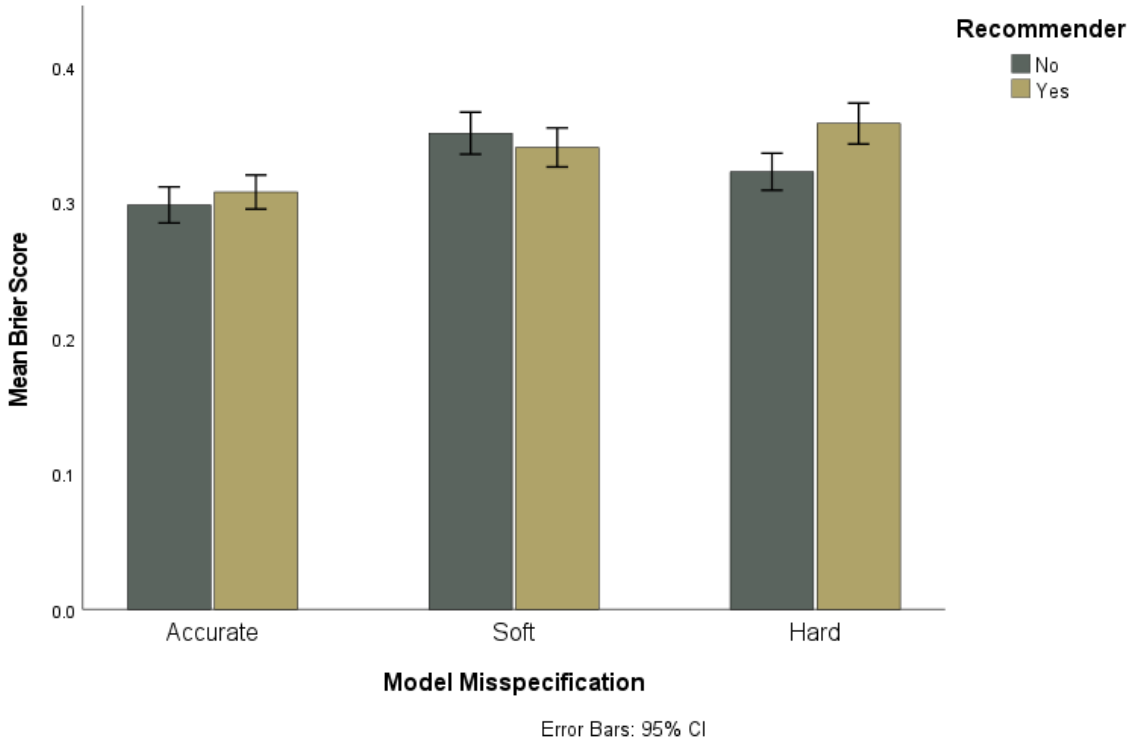

Figure 29. Mean Brier score by model misspecification and recommender conditions

For confidence judgments, none of the predicted main effects of model misspecification and recommender availability or an interaction between the two were found reliable. The only effects observed were main effect of trial ID (χ^2 (39) = 222.9, $p < 0.001$) and the interaction between trial ID and model misspecification (χ^2 (78) = 200.25, $p < 0.001$).

Table 12. Linear mixed models: Likelihood ratio tests for participants' confidence judgments in their selected route

Effect	df	ChiSq	p
Model Misspecification	2	3.520	0.172
Recommender	1	3.177	0.075
Trial ID	39	222.900	< .001
Model Misspecification × Recommender	2	1.110	0.574
Model Misspecification × Trial ID	78	200.252	< .001
Recommender × Trial ID	39	42.475	0.324
Model Misspecification × Recommender × Trial ID	78	90.178	0.163

For Brier scores, the main effect of model misspecification was significant (χ^2 (2) = 7.07, p = 0.029). The main effect of trial ID (χ^2 (39) = 650.38, p < 0.001) and the interaction between trial ID and model misspecification (χ^2 (78) = 359.94, p < 0.001) were also statistically reliable. The post-hoc contrasts for model misspecification indicate significant differences between accurate vs. hard and soft conditions:1) the accurate condition had significantly lower Brier scores compared to soft (MD=0.043; z = 5.16; p=0.01), 2) the accurate condition had lower Brier scores compared to hard (MD= 0.037; z = 17.71; p=0.03), and 3) the Brier score for the soft condition was not significantly different from hard (MD= 0.005; z = 12.56; p=0.76). Hence, participants' Brier scores were only calibrated when the system was highly accurate compared to misspecified systems. There was no difference between soft and hard conditions' Brier scores, even though they differ in accuracy.

Table 13. Linear mixed models: Likelihood ratio tests for Brier score (calibration)

Effect	df	ChiSq	p
Model Misspecification	2	7.066	0.029
Recommender	1	0.688	0.407
Trial ID	39	650.379	< .001
Model Misspecification × Recommender	2	1.753	0.416
Model Misspecification × Trial ID	78	359.940	< .001
Recommender × Trial ID	39	34.390	0.680
Model Misspecification × Recommender × Trial ID	78	80.431	0.403

3.2.6.3 Reliance and Trust in MDSS

Reliance

The reliance measure was coded as a binary variable with a score of 1 on trials when the participant's selected route belonged to the recommended set in the experimental condition and 0 on trials when it was a non-recommended route. The corresponding control conditions' reliance was also calculated based on their corresponding recommended set from the experimental condition. This estimated how likely participants in the control conditions (without any recommender) were to pick the recommended routes in experimental conditions.

The proportion of trials with reliance equal to 1 for all conditions is shown in Figure 30. The proportion of trials with reliance equal to 1 for all conditions over trial block order is shown in Figure 31. A repeated measures ANOVA using a generalized linear mixed effects model with binomial family and logit link function was used to test the effect of model misspecification, recommender, and trial block on reliance (Table 14). The trial block was used instead of trial ID or trial order as a predictor in the generalized linear mixed effects model as the large number of levels of trial ID cannot be estimated by the

model for the available data. The trial block is an ordinal variable with 4 levels, with each level comprising 10 trials in order of presentation.

The main effect of the recommender (χ^2 (1) = 12.03, $p < 0.001$) was significant, indicating a difference in reliance when the recommender was available vs. not. This was an expected effect indicating a manipulation check that participants are likely to use a recommender when it is present, even if it is misspecified. The interaction between recommender and model misspecification (χ^2 (78) = 931.76, $p < 0.001$) was also significant. Reliance is the only variable in Experiment 1 for which this predicted interaction effect was significant. The post-hoc planned contrasts for interaction effect show following differences:1) recommender conditions has significantly higher reliance compared to no-recommender control conditions (MD=0.196; z = -3.63; p<0.001), 2) the accurate recommender condition has significantly higher reliance compared to the accurate control condition (MD=0.128; z = -3.89; p<0.001), 3) the soft recommender condition has significantly higher reliance compared to the soft control condition (MD=0.065; z = -2.10; p=0.036), and 4) reliance in the hard recommender condition does not significantly differ from the hard control condition (MD=0.004; z = -0.127; p=0.899).

I predicted that the odds of selecting a recommended route (reliance) would decrease as a function of trial progression in both hard and soft misspecified conditions but that it remains relatively constant for the accurate condition. However, no evidence was found for the prediction as there were no significant trial block effects. I also expected that reliance would be lowest on recommended set in the hard condition followed by soft and that accurate would have the highest reliance among recommender conditions because of the increase in the accuracy of the preferred route set.

Table 14. Generalized linear mixed models: Likelihood ratio tests for reliance on the recommended set

Effect	df	ChiSq	p
Model Misspecification	2	4.018	0.134
Recommender	1	12.030	< .001
Trial Block	3	7.443	0.059
Model Misspecification × Recommender	2	6.816	0.033
Model Misspecification × Trial Block	6	6.952	0.325
Recommender × Trial Block	3	4.841	0.184
Model Misspecification × Recommender × Trial Block	6	8.600	0.197

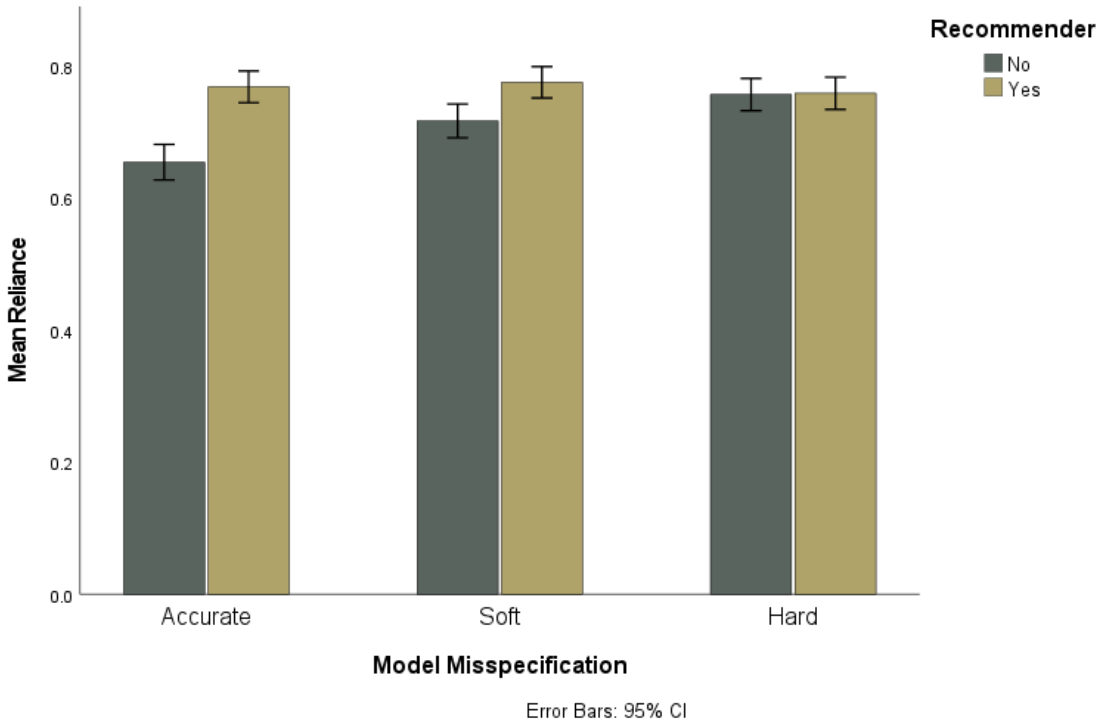

Figure 30. Mean reliance (proportion of trials when selected route belonged to recommended set) by model misspecification and recommender conditions

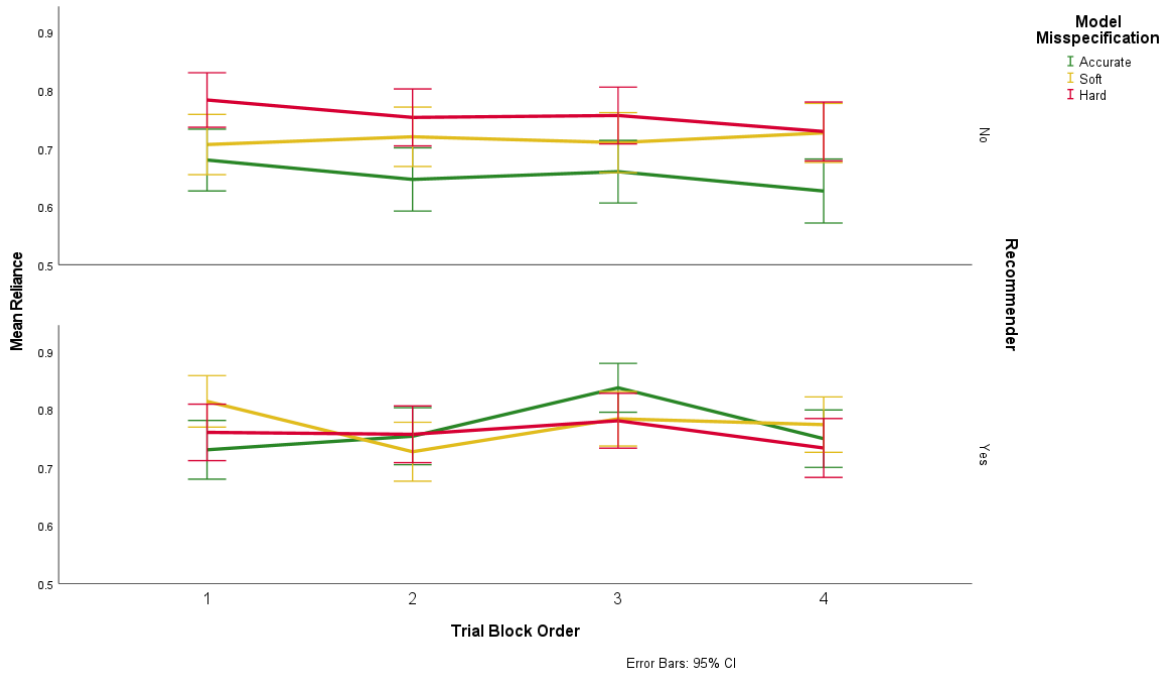

Figure 31. Mean reliance (proportion of trials when selected route belonged to recommended set) by model misspecification and recommender conditions for all trial blocks

Omnibus Trust Scores

As trust was only measured once in the experiment, a two-way ANOVA examined the effect of model misspecification and recommender availability on the omnibus trust score. The omnibus trust score was measured by taking the mean of participant responses on all 14 items of the scale (Table 7). The mean trust score plotted by condition is shown in Figure 32. No significant main effects or interactions were found (Table 15). Figure 33 shows the mean trust score for each facet of trust described before, plotted by condition. Similar to omnibus trust scores, two-way ANOVAs were conducted for each facet of trust, but no significant effects were found. Hence, there is no evidence that participants' trust in the route recommender system differs between the different conditions of the experiment.

I predicted to see a main effect of model misspecification for trust score, where participants' trust would be in the order of misspecification, with accurate having the highest trust from the participants, then soft misspecification, then hard. This is because trust in a system is proportional to the accuracy and reliability of the information provided by that system (Lee & See, 2004). I also predicted an interaction effect where trust would be better in the accurate experimental condition than in the accurate control condition. Counter-intuitively, I expected the trust in the soft and hard control conditions would be higher than in the corresponding soft and hard recommender conditions. The presence of misspecified recommenders can lead to distrust in MDSS in experimental conditions, plausibly making the trust score even worse than in control conditions. However, no statistical evidence was found for this prediction.

Table 15. Two-way ANOVA: Test of between-subject effects for omnibus trust scores

Predictor	Type III Sum of Squares	df	Mean Square	F	p
Intercept	408.330	1	408.330	2395.86	.000
Recommender	.131	1	.131	.771	.381
Model Misspecification	.160	2	.080	.470	.626
Recommender × Model Misspecification	.042	2	.021	.124	.884
Error	28.803	169	.170		

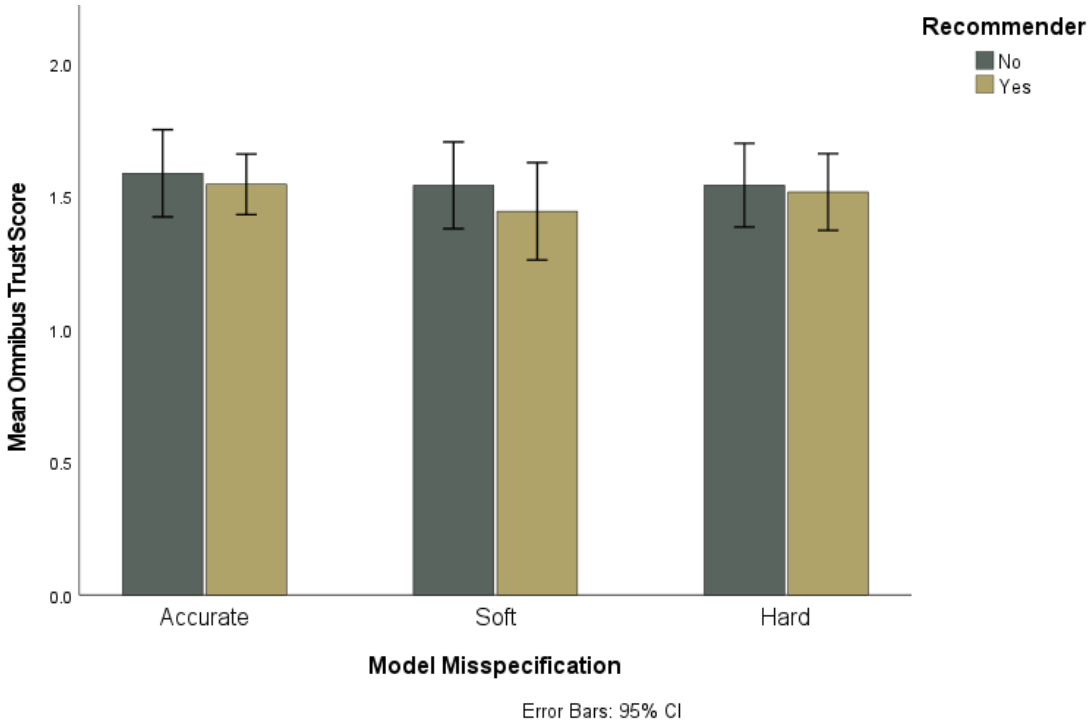

Figure 32. Mean omnibus trust score by model misspecification and recommender conditions

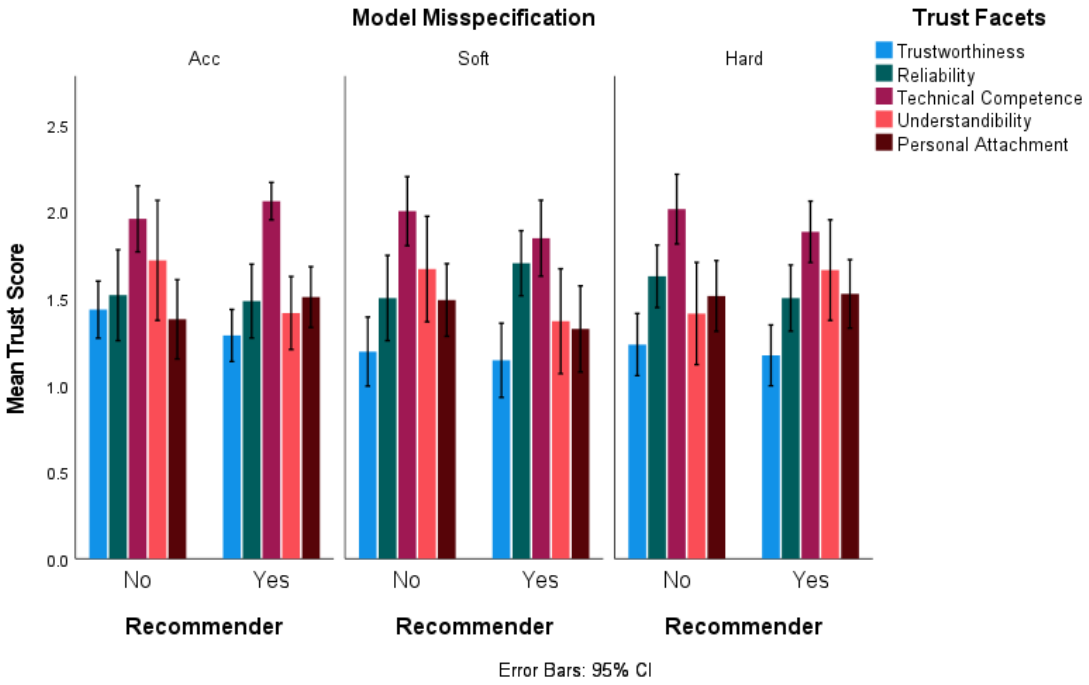

Figure 33. Mean trust score for individual trust facets measured by the trust scale by model misspecification and recommender conditions

3.2.6.4 Learning

Participants' learning was evaluated by testing their response rank across trial presentation order (Figure 34) and trial block order (Figure 35) to see if the response rank gets lower (performance improvement) with trial progression. A repeated measures ANOVA using a linear mixed-effects model was used to test the effect of recommender, model misspecification, and trial presentation order on participant response rank (Table 16). Another repeated measures ANOVA using a linear mixed-effects model was conducted to test the effect of recommender, model misspecification, and trial block order on participant response rank (Table 17). No evidence for the learning effect was found in both the repeated measures models. The main effect of the trial order, as well as the trial block, was not statistically significant nor any interaction with them were significant. This was expected to occur, as shown previously, performance across different trial IDs significantly differs across conditions. Hence, the difference in the difficulty of different scenarios might have affected the ability to detect any reliable learning effects in the task.

Table 16. Linear mixed models: Likelihood ratio tests for response rank to evaluate learning over trial presentation order

Effect	df	ChiSq	p
Model Misspecification	2	2.769	0.250
Recommender	1	6.226	0.013
Trial Order	39	45.569	0.218
Model Misspecification × Recommender	2	2.137	0.344
Model Misspecification × Trial Order	78	87.153	0.224
Recommender × Trial Order	39	32.189	0.772
Model Misspecification × Recommender × Trial Order	78	65.307	0.847

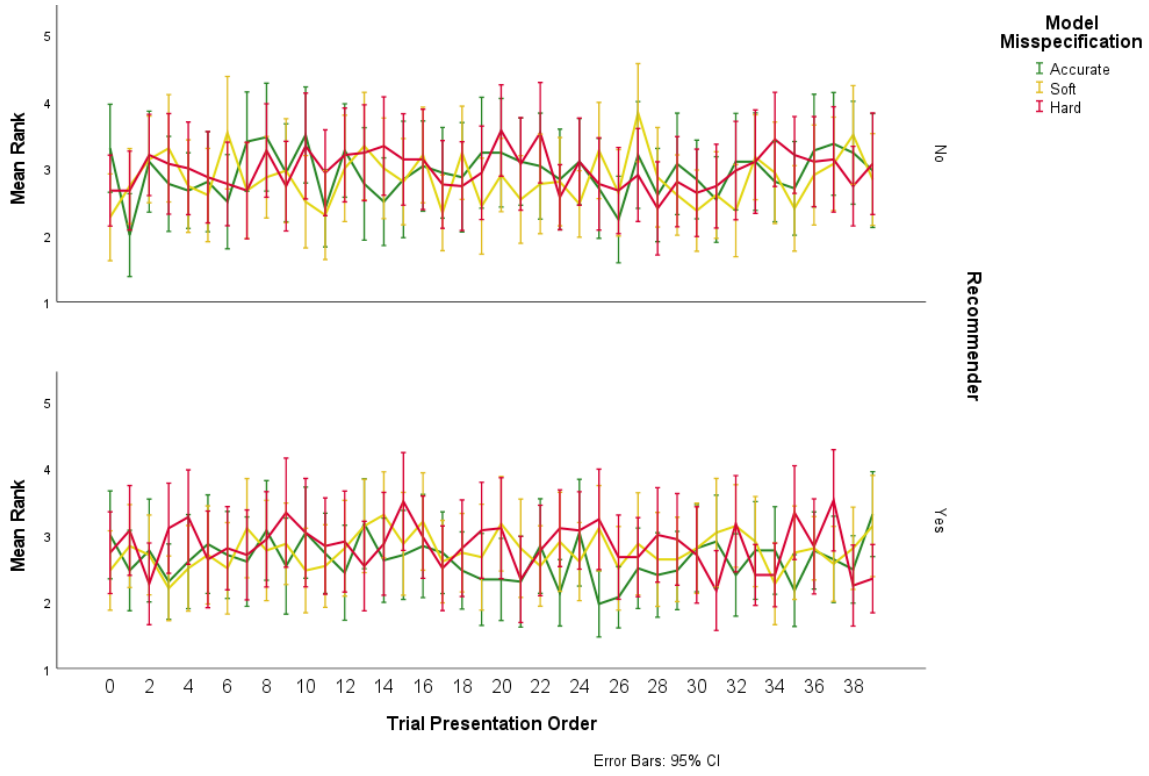

Figure 34. Mean response rank by model misspecification and recommender conditions over trial presentation order

Table 17. Linear mixed models: Likelihood ratio tests for response rank to evaluate learning over trial block order

Effect	df	ChiSq	p
Model Misspecification	2	2.779	0.249
Recommender	1	6.247	0.012
Trial Block	3	1.949	0.583
Model Misspecification × Recommender	2	2.142	0.343
Model Misspecification × Trial Block	6	3.974	0.680
Recommender × Trial Block	3	0.757	0.860
Model Misspecification × Recommender × Trial Block	6	6.071	0.415

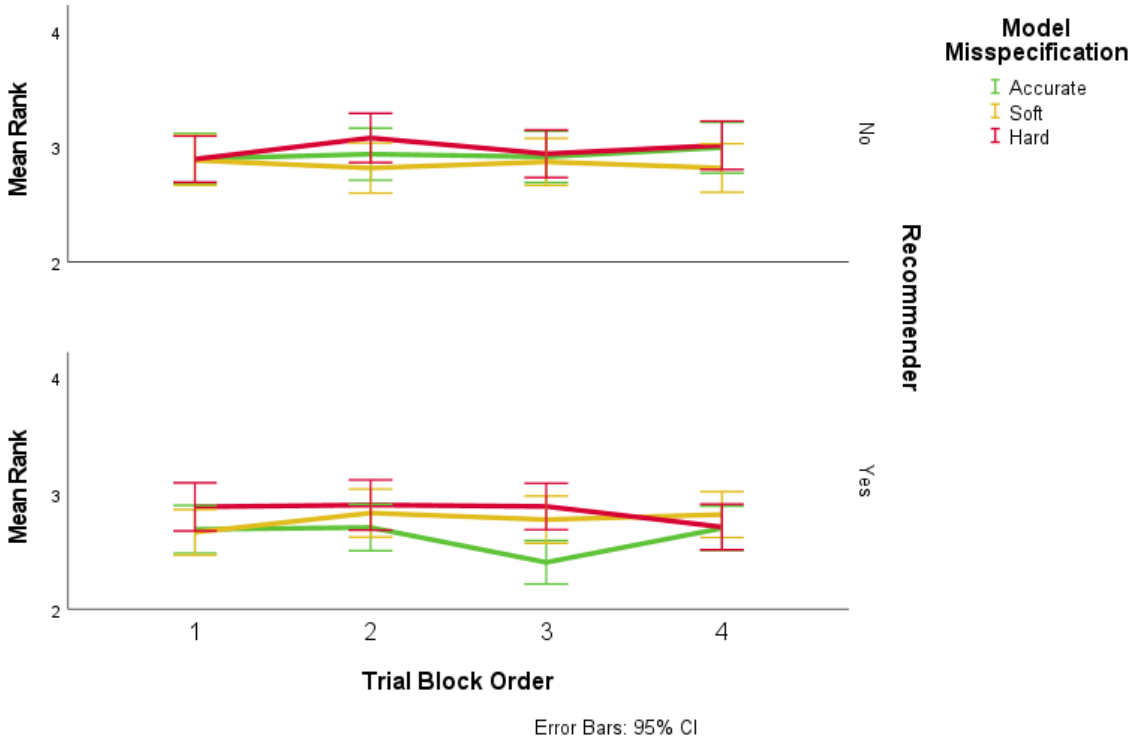

Figure 35. Mean response rank by model misspecification and recommender conditions over trial block presentation order

3.3 Experiment 2: Investigating a Model Blindness Mitigation Technique

The goal of the second study in this dissertation was to mitigate model blindness imposed on users via MDSS model misspecifications. The route recommender system from Experiment 1 was augmented by adding natural language explanations about misspecifications. The aim was to test whether XAI intervention can help calibrate decision-makers to the capabilities and limitations of an MDSS and consequently improve performance. This experiment was designed to understand how model-limited and strategy-limited can be mitigated in a complex decision-making task in an unfriendly environment.

3.3.1 Design

This experiment implements a model-blindness mitigation technique for the two misspecified conditions with MDSS-preferred recommended routes from Experiment 1: hard and soft. Both control and experimental conditions for hard and soft misspecified MDSS from Experiment 1 served as control conditions for the mitigated soft and hard experimental conditions for this experiment (Figure 36). The experiment was a 3 (Route Recommender System) x 2 (Model Misspecification) between-subjects design, as shown in Figure 36. The recommender system manipulation had three levels: (1) three MDSS-preferred recommended routes absent (Figure 13), (2) three MDSS-preferred recommended routes present (Figure 18), and (3) three MDSS-preferred recommended routes present with bias explanation (Figure 37). The model misspecification manipulation had two levels: (1) soft misspecified model and (2) hard misspecified model.

I implemented a "model blindness awareness" technique, where I mitigated model blindness by informing decision-makers about the misspecification in the model. This was done by providing a natural language explanation (adapted from XAI literature) of misspecification to the participants. The explanation message for the soft condition is shown in Figure 37.

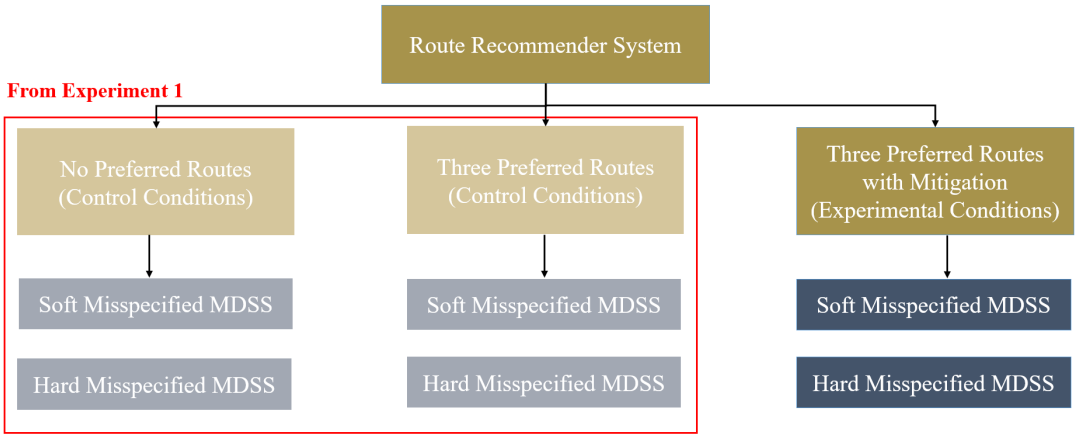

Figure 36. Design for Experiment 2

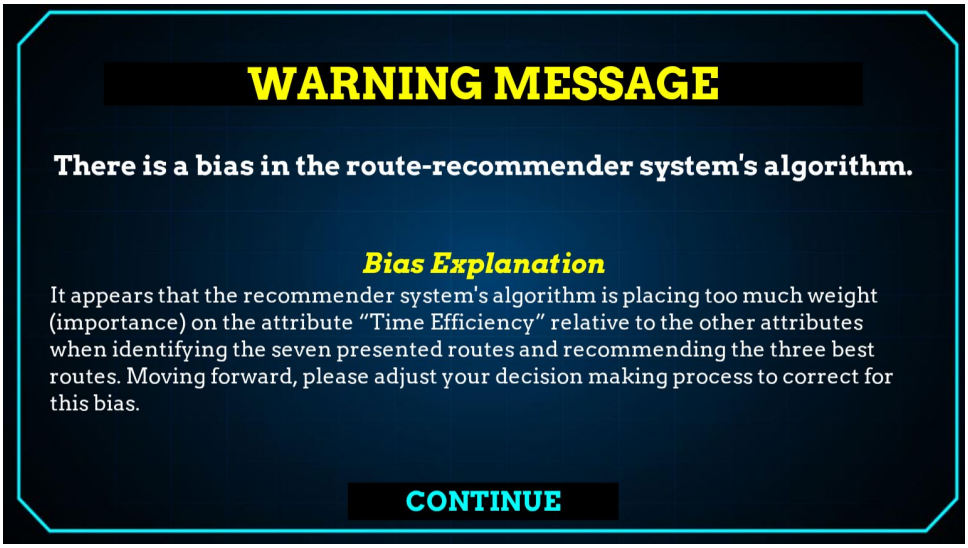

Figure 37. Bias explanation message for soft misspecified MDSS

3.3.2 Participants

62 new participants were recruited to participate in Experiment 2 explanation conditions, and all were included in the analysis. Participants' inclusion criteria, exclusion criteria, compensation, and everything else were the same as in Experiment 1. After combining data from both experiments for soft and hard model misspecification conditions,

Experiment 2 used the total data of 182 participants – at least 30 for each condition for analyses for a complete full-factorial design with a total of 6 conditions. The soft explanation condition has 32 participants.

3.3.2.1 Participant Demographics from the Pre-Study Questionnaire

The participant demographics look very similar to that of Experiment 1. It was expected because only 60-62 participants differed between both experiments' full datasets. I only have demographics questionnaire responses for 168 participants (92% participants) as the pre-study questionnaire (See APPENDIX B) was added to the study after beginning data collection. The mean age for 165 participants (3 didn't report) was 19.38 years (Range: 18-27 years). The gender distribution for participants is presented in Table 18, and their major distribution at Georgia Tech is presented in Figure 37. Computer Science was the most represented major among participants. The majority of participants (84%) had past experience playing video games (No experience- 8%, Missing Data- 8%). The majority of participants (85%) also had past experience with AI-based recommender systems (e.g., Netflix movie recommendations, Amazon product recommendations, Health and Fitness apps) in their day-to-day life (No experience- 5%, Missing data- 10%). A small number of participants (12%) had also taken a course/worked with recommender systems in the past (No course- 79%, Missing data- 10%).

Table 18. Gender Distribution of participants in Experiment 2

Gender	Percentage of Participants
Female	31%
Male	58%
Non-binary/ Non-conforming / Other	3%
Didn't report	2%
Missing Data	6%

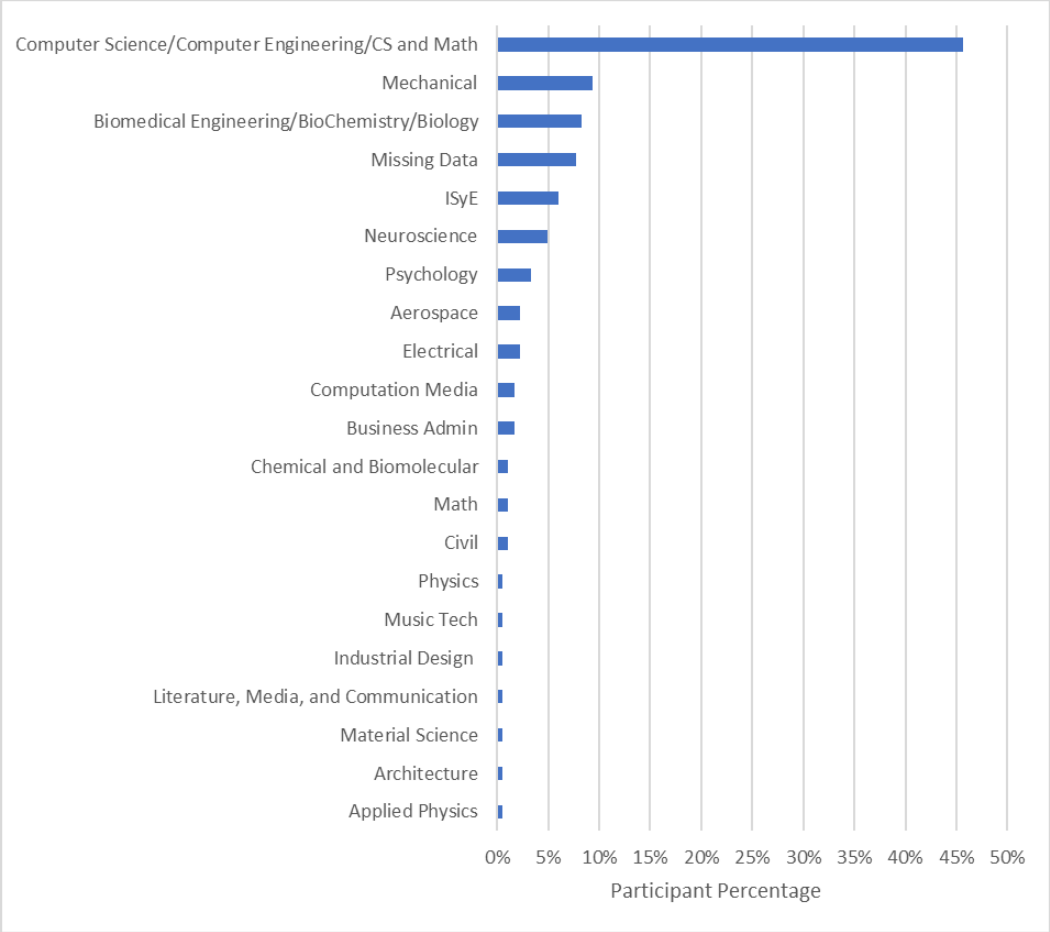

Figure 38. Major distribution of participants in Experiment 2

3.3.3 Procedure

The procedure (Figure 38) was similar to Experiment 1, and all dependent measures and analyses stayed the same. The same 40 trials of soft and hard recommender conditions were used for their respective mitigation conditions. The only difference from Experiment 1 was the presence of an explanation. The participants were reminded every five trials about the attribute overweighted by the MDSS so they could use the information to underweight those attributes in their decision-making (See APPENDIX A) for detailed participant instructions).

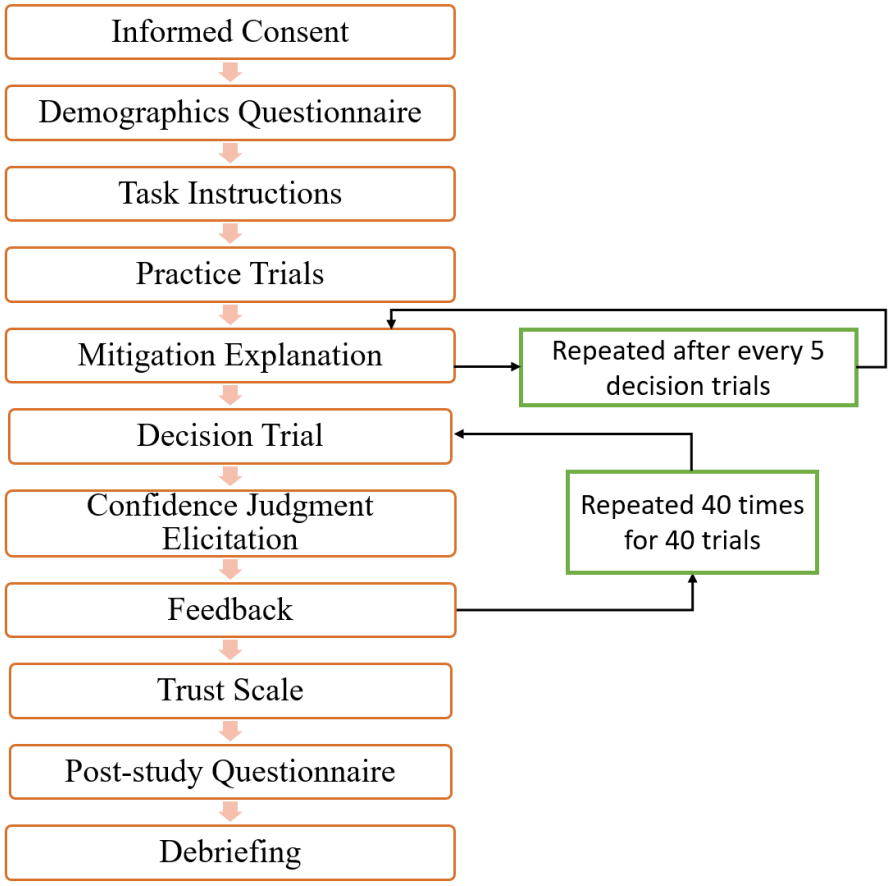

Figure 39. Procedure steps for Experiment 2

A participant in the soft mitigation condition received an explanation, as shown in Figure 37. The explanation message explicitly prompted participants to underweight the misspecified attribute – "time efficiency." A similar explanation message was shown in the hard condition, which asked participants to underweight both misspecified attributes – "time efficiency" and "obstacle avoidance." Two additional questions were added to the post-study questionnaire for Experiment 2 to get participant feedback on the presented explanation message (See APPENDIX C for the post-study questionnaire).

3.3.4 Analysis and Results

All analyses for Experiment 2 are same as Experiment 1 and organized in the same order. For all my DVs, I expected to find significant main effects of both independent variables: recommender [conditions: explanation (with explanation message and recommender preferred 3 routes), yes (with recommender preferred 3 routes), no (without recommender preferred 3 routes)] and model misspecification (conditions: soft and hard). I also expected to see a significant interaction between model misspecification and recommender. However, little to no evidence was found for the main effect of model misspecification, the main effect of recommender, and the interaction between recommender and model misspecification for most of my DVs. The following sub-sections discuss the results organized by DVs.

3.3.4.1 Performance Measures

Response rank

Figure 40 shows the mean rank of participant route choices by each condition in Experiment 2. A repeated measures ANOVA using a linear mixed-effects model was used to test the effect of recommender and model misspecification on participant decision choices using response rank as a continuous variable. The trial ID (scenarios) variable was also added as a predictor in the model. The ANOVA summary table is shown in Table 19. The main effect of trial ID (χ^2 (39) = 971.36, $p < 0.001$) was significant, indicating a difference in the difficulty level of different trials affecting performance. The interaction between trial ID and model misspecification (χ^2 (39) = 462.87, $p < 0.001$) was also significant, indicating that performance on different trials varied significantly across model misspecification conditions. The interaction between trial ID and recommender (χ^2 (78) = 113.88, $p = 0.005$) was also significant, indicating that performance on different trials varied significantly across recommender conditions. The performance for all trial IDs across conditions is shown in Figure 41.

Table 19. Linear mixed models: Likelihood ratio tests for response rank

Effect	df	ChiSq	p
Model Misspecification	1	0.186	0.666
Recommender	2	3.509	0.173
Trial ID	39	971.360	< .001
Model Misspecification × Recommender	2	1.808	0.405
Model Misspecification × Trial ID	39	462.868	< .001
Recommender × Trial ID	78	113.880	0.005
Model Misspecification × Recommender × Trial ID	78	90.268	0.162

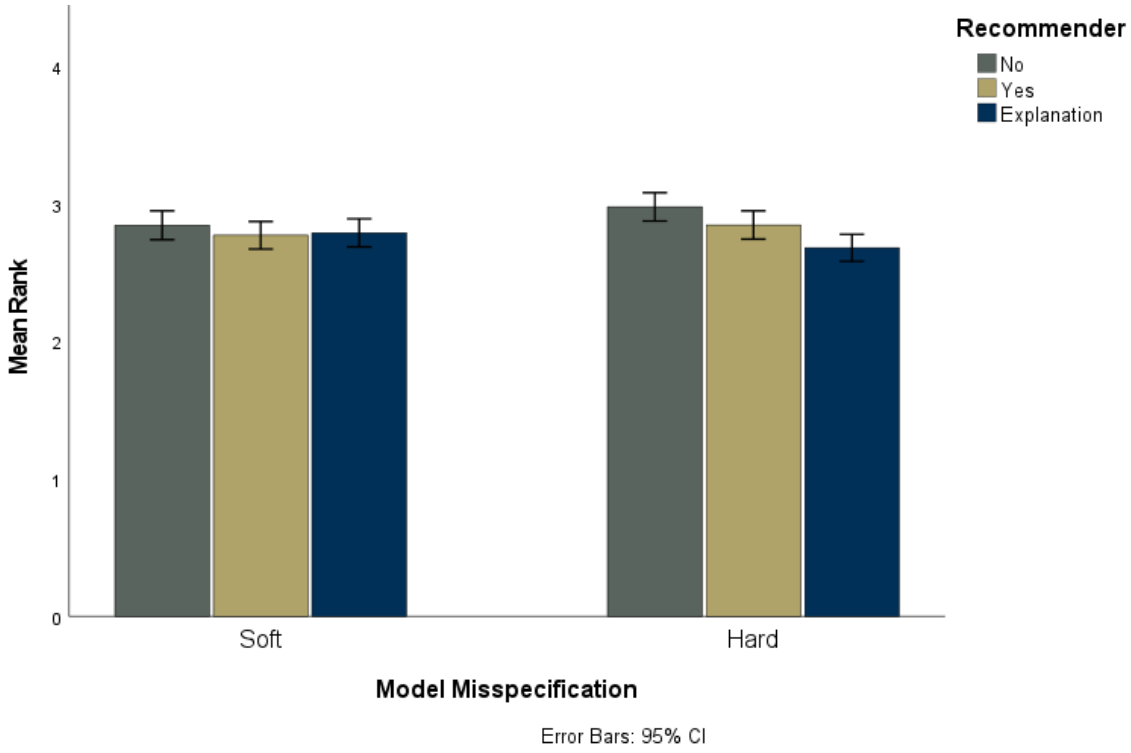

Figure 40. Mean response rank (out of 7) of selected route by model misspecification and recommender conditions

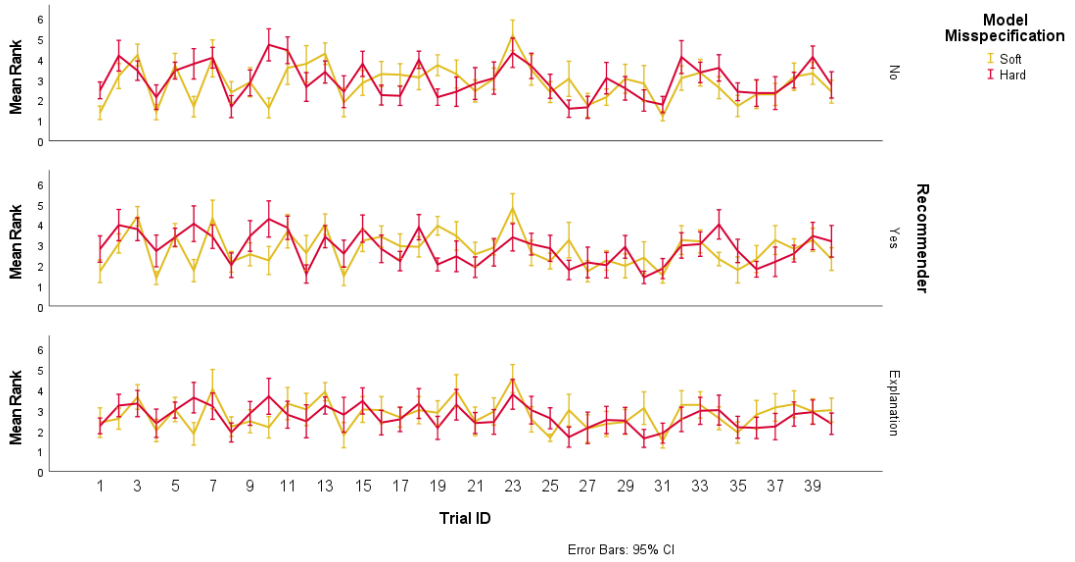

Figure 41. Mean response rank (out of 7) of the selected route by model misspecification and recommender conditions for all Trial IDs

I predicted a main effect of recommender where participants in explanation conditions would perform better (lower mean rank) than control conditions (without explanation – both yes and no recommender conditions). I also predicted a main effect of model misspecification, where participants in the soft condition would perform better than the hard condition, because the soft condition requires the adjustment (underweighting) of only one attribute and adjusting (underweighting) for two attributes simultaneously was required in the hard condition. I also predicted an interaction between recommender and model misspecification. There were no significant main effects of model misspecification or recommender, nor the interaction between model misspecification and recommender. Thus, the data provided no statistical evidence for the main effects or interaction predictions for mean route ranks.

Outcome

All the predictions for outcome DV were the same as the response rank DV. The proportion of trials with the outcome equal to 1 for all conditions of Experiment 2 is shown in Figure 42. A repeated measures ANOVA using a generalized linear mixed effects model with binomial family and logit link function was used to test the effect of model misspecification, recommender availability, and trial block on the outcome. No significant effects were found (Table 20). Hence, no evidence was found for the predicted main effect of model misspecification, recommender, or their interaction. No significant difference in participants' ability to select the best outcome changed with the misspecification or recommender level.

Table 20. Generalized linear mixed models: Likelihood ratio tests for outcome

Effect	df	ChiSq	p
Model Misspecification	1	0.605	0.437
Recommender	2	3.494	0.174
Trial block	3	1.267	0.737
Model Misspecification × Recommender	2	1.947	0.378
Model Misspecification × Trial block	3	2.129	0.546
Recommender × Trial block	6	3.001	0.809
Model Misspecification × Recommender × Trial block	6	5.963	0.427

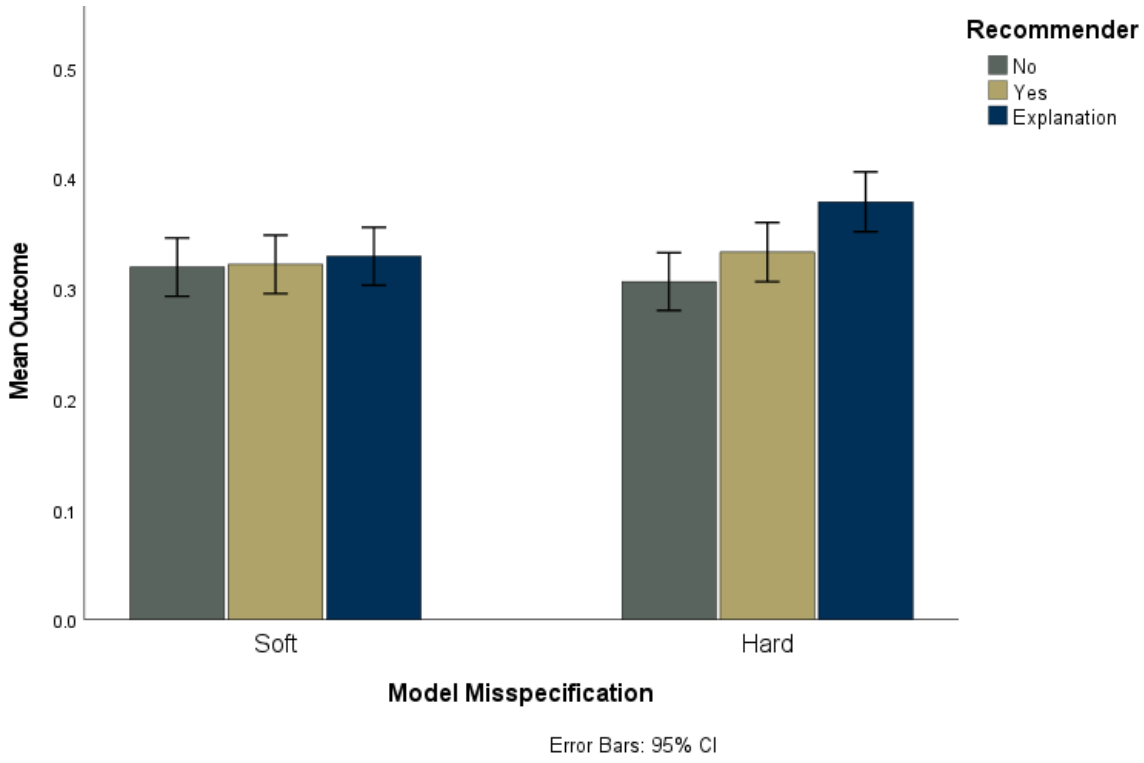

Figure 42. Mean outcome (proportion of trials with the best route selected) by model misspecification and recommender conditions

Local utility loss

The local utility loss (Figure 41) is again a performance measure with the same predictions as response rank and outcome. The lower utility loss is considered better. A repeated measures ANOVA using a linear mixed-effects model was used to test the effect of recommender and model misspecification on participant decision choices using local utility loss as a continuous variable. The trial ID was also included as a predictor in the model to test if the local utility loss varies significantly across different trials. The ANOVA summary table is shown in Table 21. The significant effects are similar to the response rank effects presented before. The main effect of trial ID (χ^2 (39) = 1109.40, $p < 0.001$) was found significant, indicating a difference in trial difficulty level affecting performance. The interaction between trial ID and model misspecification (χ^2 (39) = 844.02, $p < 0.001$) was also significant, indicating that performance on different trials varied significantly across misspecification conditions. Some trials were more difficult than others at some misspecification levels.

Table 21. Linear mixed models: Likelihood ratio tests for local utility loss of selected route

Effect	df	ChiSq	p
Model Misspecification	1	1.866	0.172
Recommender	2	2.514	0.285
Trial ID	39	1109.398	< .001
Model Misspecification × Recommender	2	1.345	0.510
Model Misspecification × Trial ID	39	844.021	< .001
Recommender × Trial ID	78	97.054	0.071
Model Misspecification × Recommender × Trial ID	78	89.559	0.175

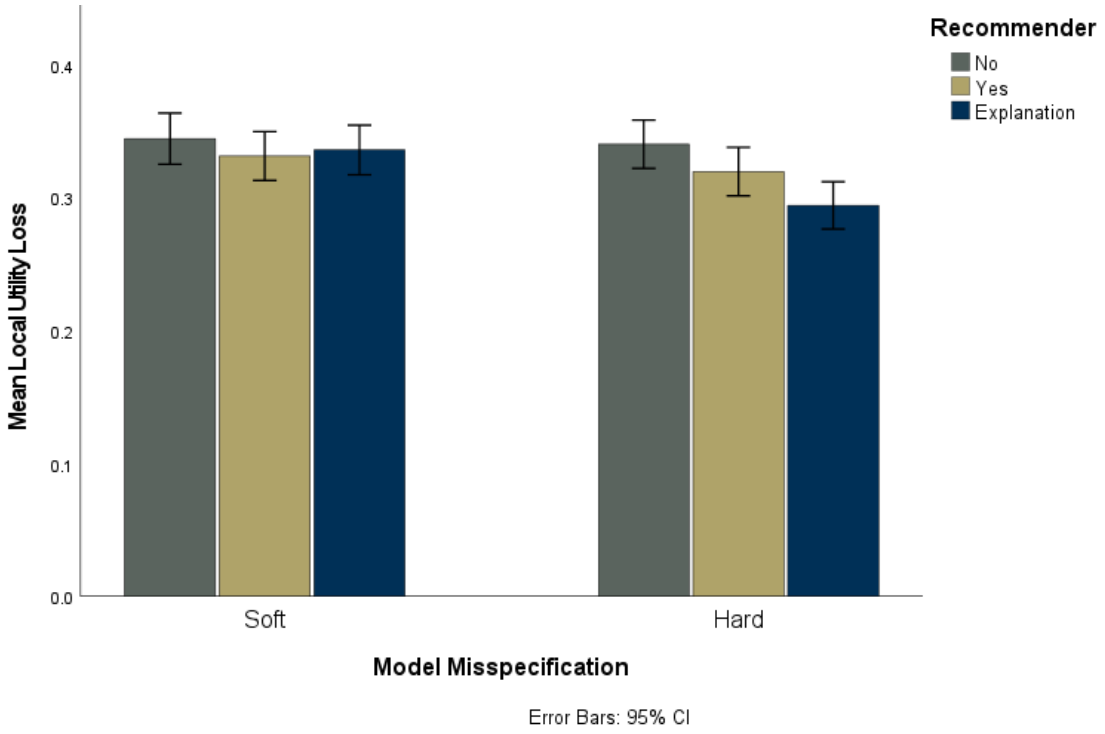

Figure 43. Mean local utility loss of selected route by model misspecification and recommender conditions

Global utility loss

The global utility loss (Figure 44) is again a performance measure like local utility loss, with a lower value being considered better. However, global utility loss evaluates participants' route choice in comparison to the true global best route generated by the true equal weights model. It compares route choices to the true world for MDSS instead of just seven presented alternatives. As it represents the true world, I expected it to follow the order of misspecification (main effect) for all levels of the recommender. A repeated measures ANOVA using a linear mixed-effects model was used to test the effect of recommender and model misspecification on participant decision choices using global utility loss as a continuous variable. The trial ID was also included as a predictor in the

model to test if the global utility loss varies significantly across different trials. The ANOVA summary table is shown in Table 22.

Participants' global utility loss for their selected routes differed significantly between model misspecification conditions (χ^2 (2) = 99.27, p < .001). The soft condition had significantly lower global utility loss compared to the hard (MD= 0.032). This main effect of model misspecification only serves as a manipulation check for the experiment that there were significant differences in the accuracy of the MDSS at two misspecification levels that were also reflected in participant performance. The main effect of trial ID (χ^2 (39) = 1814.68, p < 0.001) was also significant, indicating a difference in their difficulty level affecting performance. The interaction between trial ID and model misspecification (χ^2 (39) =481, p < 0.001) was also significant, indicating that performance on different trials varied significantly across misspecification levels. Some trials were more difficult than others at some misspecification levels.

Table 22. Linear mixed models: Likelihood ratio tests for global utility loss of selected route

Effect	df	ChiSq	p
Model Misspecification	1	99.267	< .001
Recommender	2	2.229	0.328
Trial ID	39	1814.676	< .001
Model Misspecification × Recommender	2	2.083	0.353
Model Misspecification × Trial ID	39	480.997	< .001
Recommender × Trial ID	78	98.418	0.059
Model Misspecification × Recommender × Trial ID	78	93.607	0.110

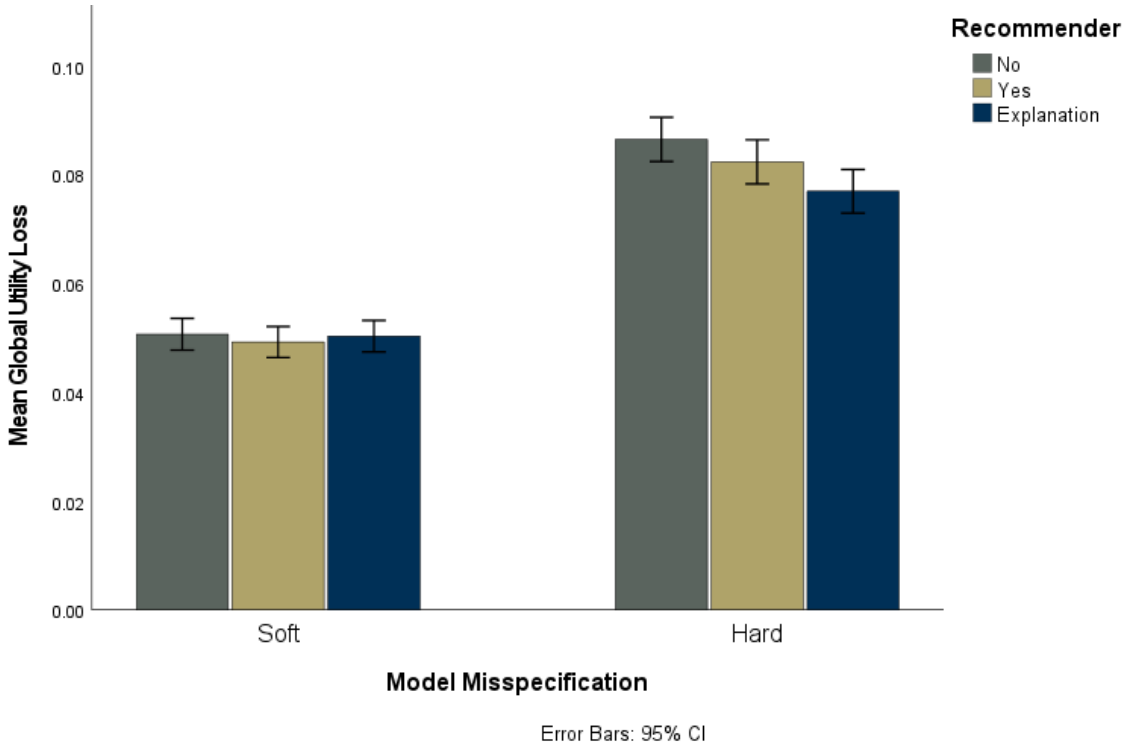

Figure 44. Mean global utility loss of selected route by model misspecification and recommender conditions

3.3.4.2 Confidence Judgement and Brier Scores

Similar to Experiment 1, I expected for Experiment 2 that confidence and calibration would be correlated with performance measures. I also expected a better calibration (lower BS) to model misspecification levels in explanation conditions compared to corresponding control conditions (main effect of recommender) because direct information about misspecification was presented to participants via explanation.

Two repeated measures ANOVA using a linear mixed-effects model were conducted to test the effect of recommender and model misspecification on participant confidence judgments (Figure 45) and Brier scores (Figure 46). The trial ID was also

included as a predictor in both models to test whether confidence judgments or Brier scores change significantly across different trials. The ANOVA summary tables are shown in Table 23 and Table 24.

For confidence judgments, none of the predicted main effects of model misspecification and recommender or their interaction were found reliable. The only effects observed were the main effect of trial ID (χ^2 (39) = 259.94, $p < 0.001$) and the interaction between trial ID and model misspecification (χ^2 (39) = 114.59, $p < 0.001$).

For Brier scores, again, the predicted main effects of model misspecification and recommender or their interaction were not statistically significant. The main effect of trial ID (χ^2 (39) = 656.50, $p < 0.001$) and the interaction between trial ID and model misspecification (χ^2 (39) = 190.48, $p < 0.001$) were significant. The three-way interaction between trial ID, model misspecification, and recommender (χ^2 (78) = 132.90, $p < 0.001$) was also statistically significant.

Table 23. Linear mixed models: Likelihood ratio tests for participants' confidence judgments in their selected route

Effect	df	ChiSq	p
Model Misspecification	1	1.033	0.309
Recommender	2	3.333	0.189
Trial ID	39	259.939	< .001
Model Misspecification × Recommender	2	1.807	0.405
Model Misspecification × Trial ID	39	114.587	< .001
Recommender × Trial ID	78	76.791	0.517
Model Misspecification × Recommender × Trial ID	78	88.151	0.202

Table 24. Linear mixed models: Likelihood ratio tests for Brier score (calibration)

Effect	df	ChiSq	p
Model Misspecification	1	0.024	0.877
Recommender	2	5.877	0.053
Trial ID	39	656.497	< .001
Model Misspecification × Recommender	2	1.637	0.441
Model Misspecification × Trial ID	39	190.482	< .001
Recommender × Trial ID	78	92.018	0.133
Model Misspecification × Recommender × Trial ID	78	132.904	< .001

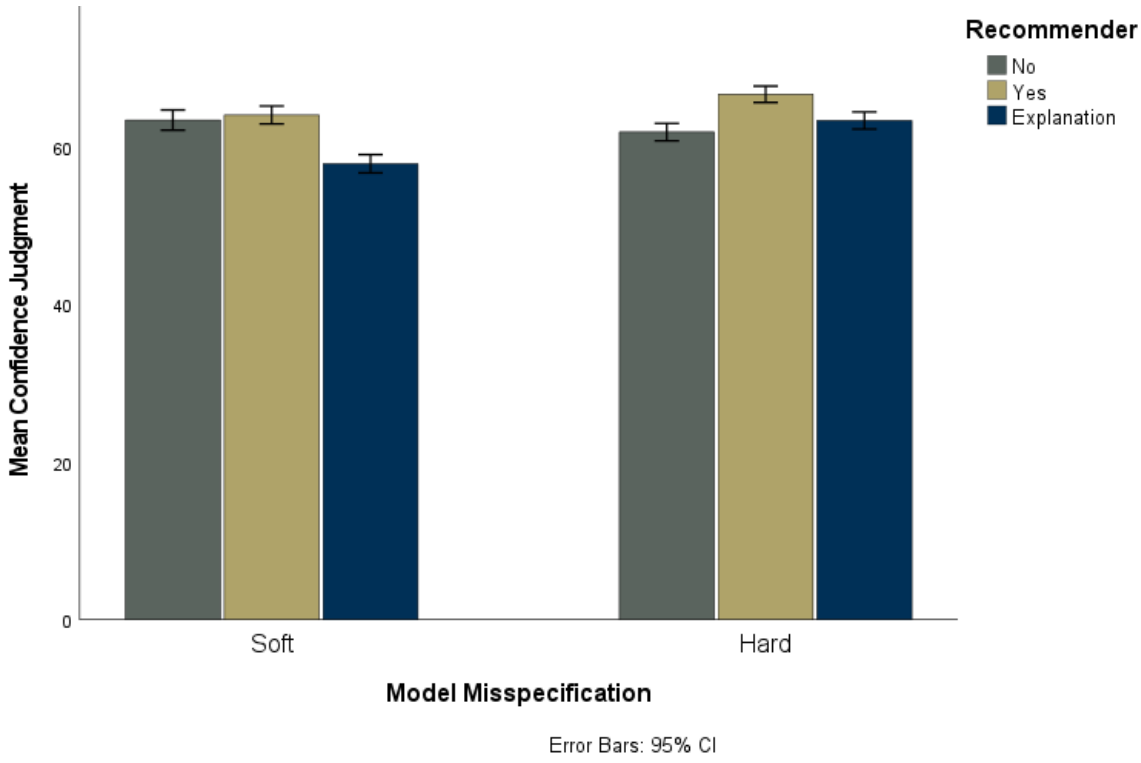

Figure 45. Mean confidence judgment score by model misspecification and recommender conditions

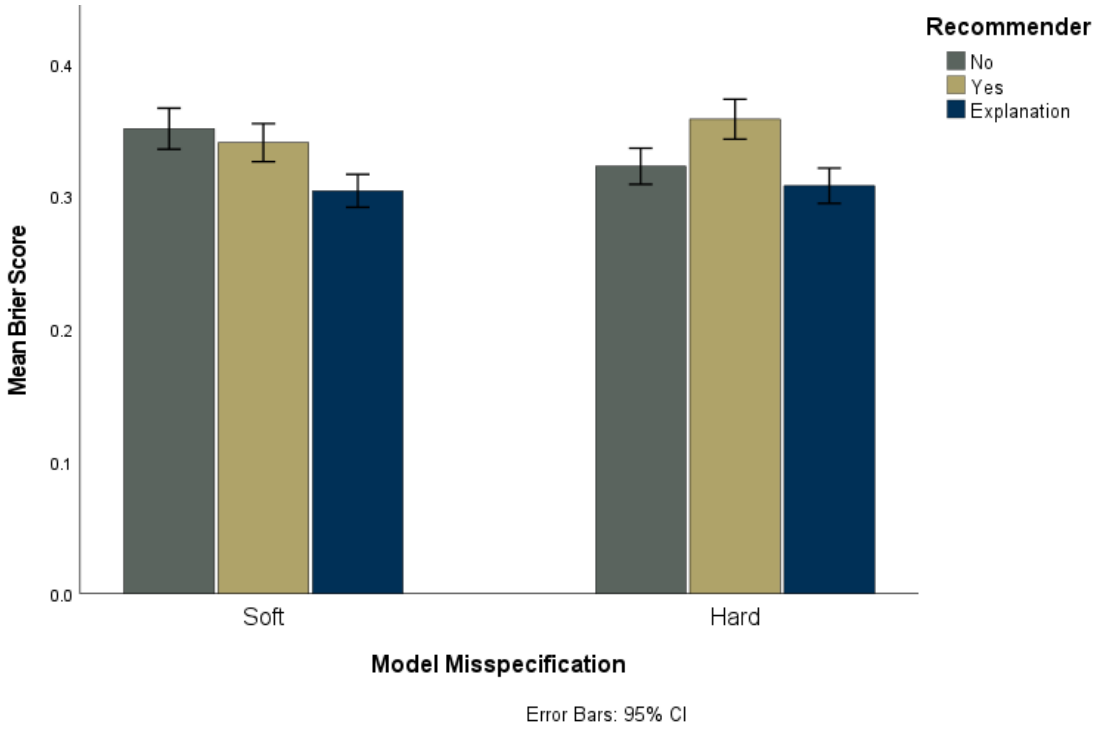

Figure 46. Mean Brier score by model misspecification and recommender conditions

3.3.4.3 Reliance and Trust in MDSS

For both reliance and trust measures, I predicted that participants' trust and reliance would worsen in the explanation conditions compared to their control conditions as the explanation explicitly informs participants about the error in the system.

Reliance

The proportion of trials with reliance equal to 1 for all conditions is shown in Figure 47. The proportion of trials with reliance equal to 1 for all conditions over trial block order is shown in Figure 48. A repeated measures ANOVA using a generalized linear mixed

effects model with binomial family and logit link function was used to test the effect of model misspecification, recommender, and trial block on reliance (Table 25).

The main effect of the recommender (χ^2 (2) = 8.35, p= 0.015) was significant, indicating a difference in reliance when the recommender was absent, present, or present with an explanation. The post-hoc contrasts show the following differences:1) recommender conditions has significantly higher reliance compared to the explanation conditions (MD=0.067; z = 2.92; p=0.003), 2) reliance in the explanation conditions does not significantly differ from its no recommender controls (MD=0.-33; z = -1.376; p=0.169), and 3) reliance in the recommender conditions does not significantly differ from its no recommender controls (MD=0.034.; z = 1.538; p=0.124). Hence, participants' reliance in the explanation conditions does not differ from that of participants that had no recommended set. Hence, providing explicit explanations does reduce reliance on a misspecified system even though no performance changes were detected between misspecification conditions.

The three-way interaction between model misspecification, recommender, and the trial block was also significant (χ^2 (6) = 16.21, p= 0.013). The interaction effect patterns are illustrated in Figure 48. No consistent patterns can be observed from this interaction effect. The performance across trial blocks at different levels of recommender varies with participants in the hard misspecification condition performing worse than soft for some blocks, the pattern reverses for other blocks, and for some blocks, the performance level stays the same for both levels of misspecification.

Table 25. Generalized linear mixed models: Likelihood ratio tests for reliance on the recommended set

Effect	df	ChiSq	p
Model Misspecification	1	0.298	0.585
Recommender	2	8.359	0.015
Trial block	3	4.178	0.243
Model Misspecification × Recommender	2	1.846	0.397
Model Misspecification × Trial block	3	1.247	0.742
Recommender × Trial block	6	2.987	0.810
Model Misspecification × Recommender × Trial block	6	16.210	0.013

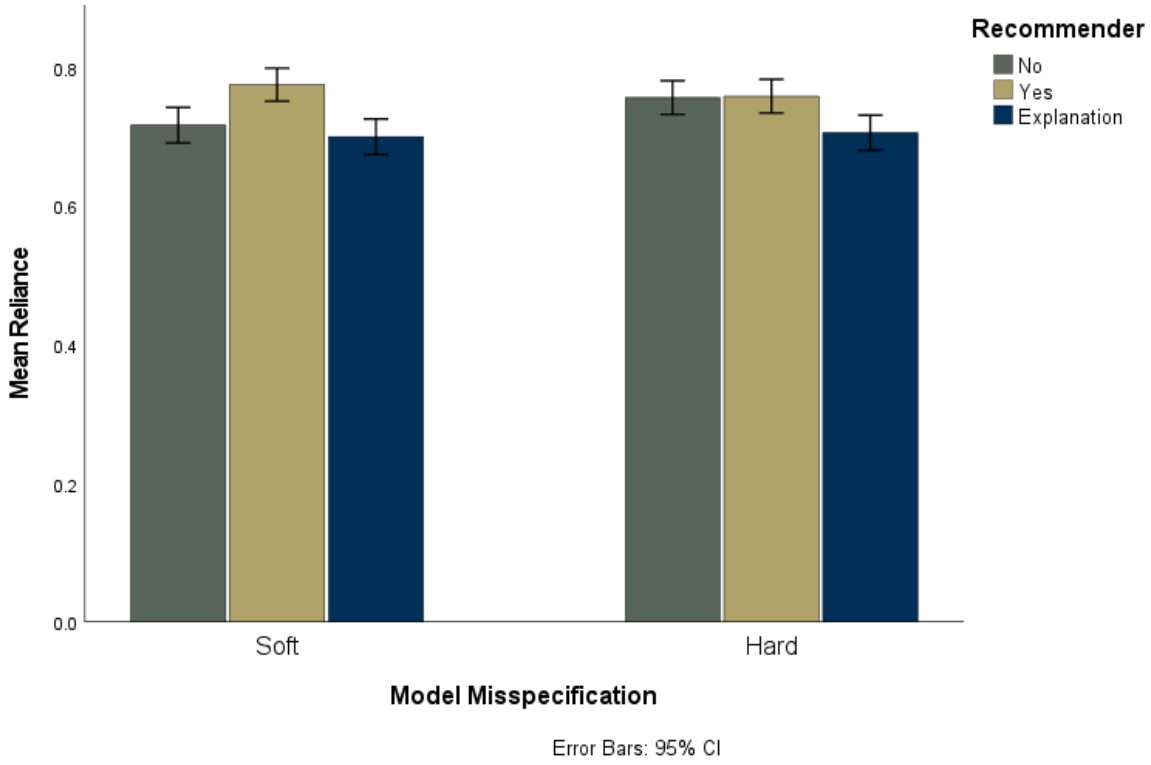

Figure 47. Mean reliance (proportion of trials when selected route belonged to recommended set) by model misspecification and recommender conditions

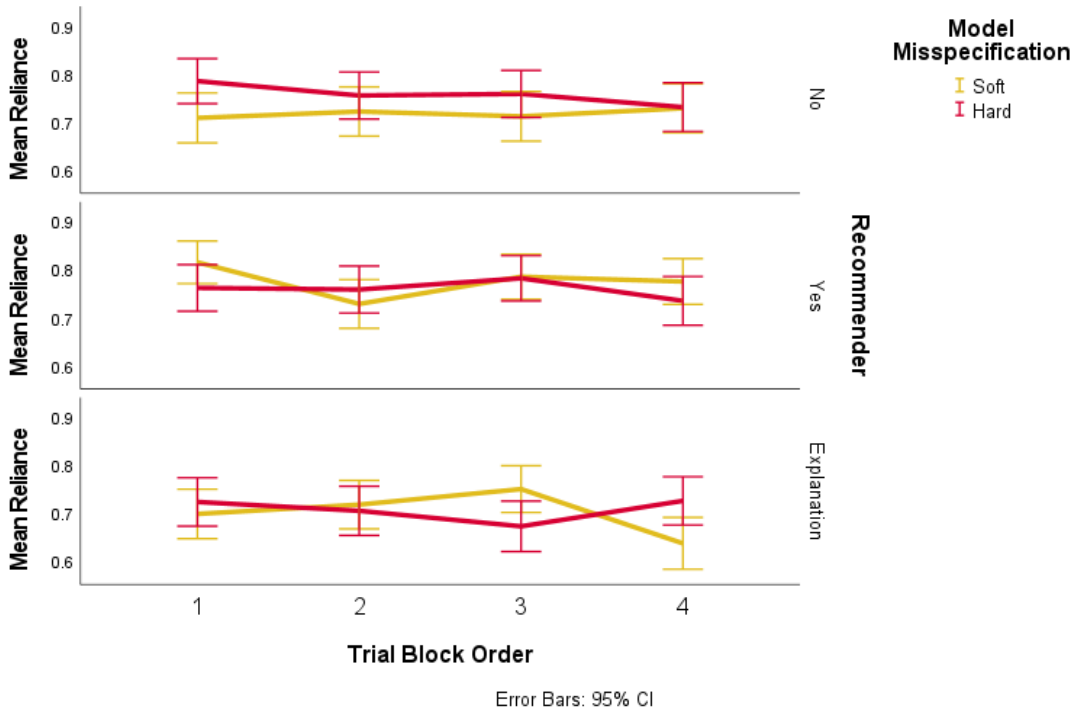

Figure 48. Mean reliance (proportion of trials when selected route belonged to recommended set) by model misspecification and recommender conditions for all trial blocks

Omnibus Trust Scores

As trust was only measured once in the experiment, a two-way ANOVA was conducted that examined the effect of model misspecification and recommender availability on the omnibus trust score. The omnibus trust score was measured by taking the mean of participant responses on all 14 items of the scale (Table 7). The mean trust score by conditions is shown in Figure 49. Figure 50 shows the mean trust score for each facet of trust by conditions. Similar to omnibus trust scores, two-way ANOVAs were conducted for each facet of trust. The ANOVA summary table for omnibus trust scores is shown in Table 26.

For the omnibus trust score, the main effect of recommender was found to be significant (F(2)= 4.33,p=0.015). The post-hoc Tukey's HSD test shows that trust in the explanation condition was significantly lower than trust in the no recommender (control) condition (MD: .225, p= 0.011). A similar main effect of recommender and post-hoc comparison results was found for trust scores for the scale's trustworthiness facet (F(2)= 3.045, p=0.050; MD=.205, p= 0.04), technical competence facet (F(2)= 4.33, p=0.015; MD= .305, p= 0.013), and personal attachment facet (F(2)= 3.86,p=0.023; MD= .275 p= 0.017). The trust score between conditions was not significantly different for the reliability and understandability facets of the scale.

Table 26. Two-way ANOVA: Test of between-subject effects for omnibus trust scores

Predictor	Type III Sum of Squares	df	Mean Square	F	p
Intercept	415.289	1	415.289	2325.49	.000
Recommender	1.549	2	.774	4.336	.015*
Model Misspecification	.122	1	.122	.682	.410
Recommender × Model Misspecification	.130	2	.065	.363	.696
Error	30.716	172	.179		

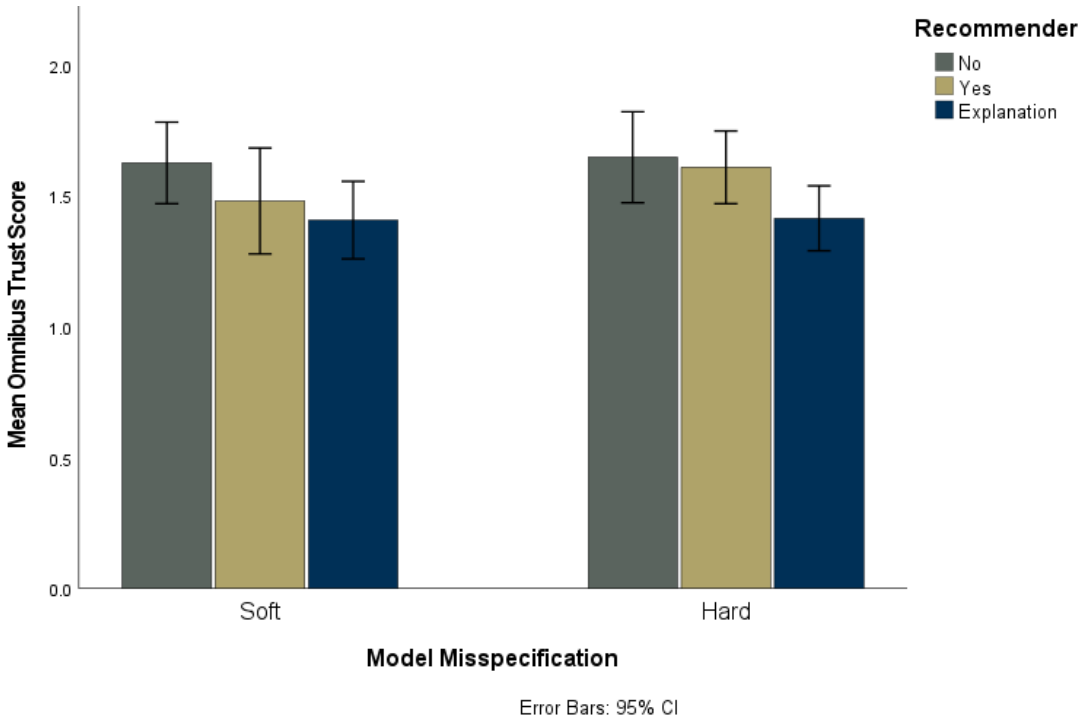

Figure 49. Mean omnibus trust score by model misspecification and recommender conditions

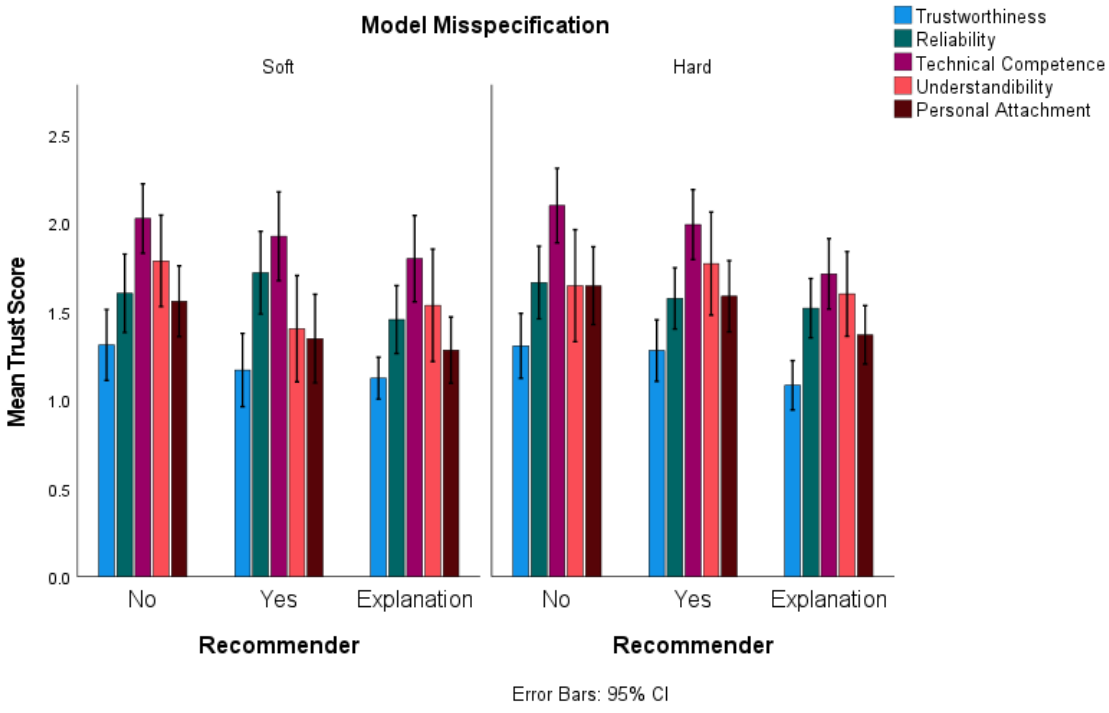

Figure 50. Mean trust score for individual trust facets by model misspecification and recommender conditions

3.3.4.4 Learning

Participants' learning was evaluated by testing their response rank across trial presentation order (Figure 51) and trial block order (Figure 52) to see if the response rank gets lower (performance improvement) with trial progression. A repeated measures ANOVA using a linear mixed-effects model was used to test the effect of recommender, model misspecification, and trial presentation order on participant response rank (Table 27). Another repeated measures ANOVA using a linear mixed-effects model was conducted to test the effect of recommender, model misspecification, and trial block order on participant response rank (Table 28). No evidence for the learning effect was found in both the repeated measures models. The main effects of the trial order, as well as the trial block, were not statistically significant. The trial block and recommender interaction was significant (χ^2 (6) = 13.07, p= 0.042). However, there was no clear performance improvement (decreasing rank) pattern with trial block progression. Hence, it is difficult to draw any specific conclusions about this interaction effect using Figure 52.

Table 27. Linear mixed models: Likelihood ratio tests for response rank to evaluate learning over trial presentation order

Effect	df	ChiSq	p
Model Misspecification	1	0.173	0.678
Recommender	2	3.486	0.175
Trial order	39	38.534	0.491
Model Misspecification × Recommender	2	1.789	0.409
Model Misspecification × Trial order	39	54.570	0.050
Recommender × Trial order	78	87.870	0.208
Model Misspecification × Recommender × Trial order	78	61.635	0.913

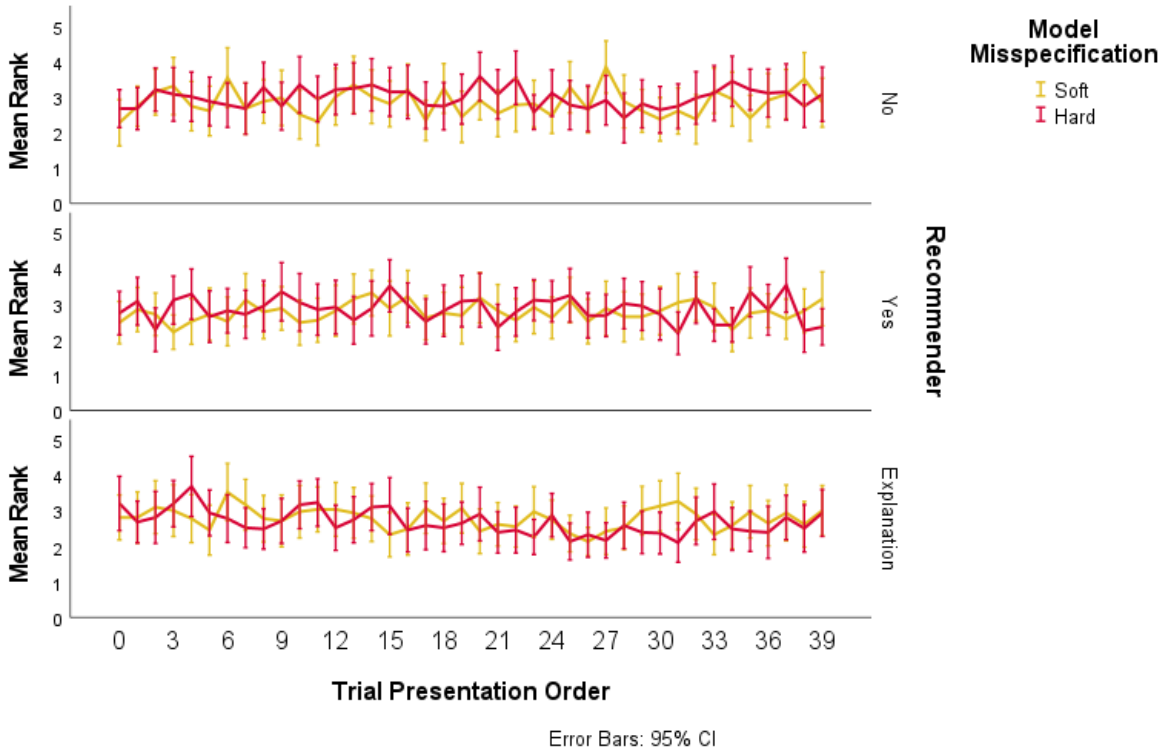

Figure 51. Mean response rank by model misspecification and recommender conditions over trial presentation order

Table 28. Linear mixed models: Likelihood ratio tests for response rank to evaluate learning over trial block order

Effect	df	ChiSq	p
Model Misspecification	1	0.174	0.677
Recommender	2	3.489	0.175
Trial block	3	6.243	0.100
Model Misspecification × Recommender	2	1.789	0.409
Model Misspecification × Trial block	3	2.168	0.538
Recommender × Trial block	6	13.069	0.042
Model Misspecification × Recommender × Trial block	6	4.514	0.607

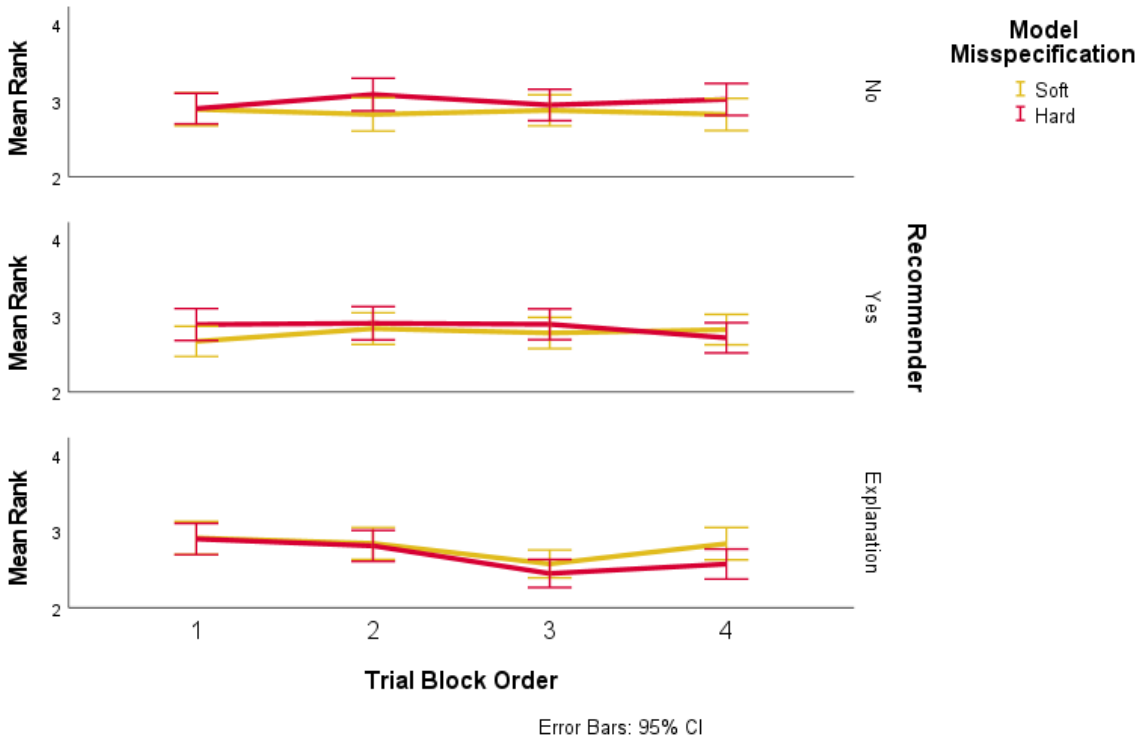

Figure 52. Mean response rank by model misspecification and recommender conditions over trial block presentation order

3.4 Participant's and Conditions' Best-fit Decision Strategy

The Bayesian approach was used to assess the fit of each strategy (from Table 1) to the participant's choice profiles— the best-fit strategy was the one that was most likely given the participant's pattern of route selections. It was interesting to look at whether participants followed more compensatory strategies accounting for tradeoffs, like EQ weights, conjunctive, and so on, or do they chose to follow more non-compensatory heuristic strategies, like lexicographic. This strategy-fitting helped investigate the role of decision strategies on performance, as discussed in the proposed model blindness confluence model (Figure 6), and provided insights about when strategy-limited performance manifests under a model-limited and context-limited decision task. The strategies were estimated in three ways: (1) for each experimental condition (collapsed

across all participants in that condition), (2) for each participant (collapsed across all trials), and (3) for each participant for every trial block. The strategies were determined using Bayesian model fitting explained in detail later in this section.

The implemented Bayesian analyses take the attribute-weighting properties of different decision strategies into account when deriving predictions concerning the probability of obtaining a particular route choice under a particular attribute-weighting scheme (decision strategy). The attribute weights for all the strategies are shown in Table 29. The first step in the analyses was operationalizing the attribute weightings for each decision strategy based on the decision task. The strategies were operationalized for the estimation processes as follows:

- A ***weighted additive (WADD) strategy*** applies true weights to the attributes and selects the option with the highest utility. In the case of the route recommender system, WADD will be ***equal weights*** for accurate conditions, ***soft corrective weights*** for soft conditions, and ***hard corrective weights*** for hard conditions. The corrective weights are the approximate weights that should be used to overcome the misspecification imposed by the model via ignoring misspecified attributes. The corrective weights strategies set misspecified attributes equal to 0 because the task ecology validation analysis on random samples of simulated responses before data collection showed that these attributes should be weighted approximately 0 to improve performance on any random sample of route selections (Table 6). Hence, all three WADDs were included in the Bayesian analyses; their weights are shown in Table 29, which were used to calculate the utility of each route.

- A ***lexicographic strategy*** is a heuristic strategy that was implemented by operationalizing it as giving weight to only one attribute out of all six presented for all seven routes, and the route choice from the lexicographic strategy should be the route with the highest value on that attribute. Table 29 shows six different lexical strategies (Lex1-Lex6) as each of them is operationalized by giving a weight "1" to only one attribute out of 6 and weighting all others as "0". The utility for all seven routes was calculated separately by each lexicographic strategy.
- A ***MaxiMax strategy*** picks the alternative that maximizes the probability of the best-case outcome. Hence, it was operationalized as taking the maximum of three attributes that were in the same direction (positively correlated with each other) to get the utility of a route. The attributes were time efficiency, obstacle avoidance, and hazard avoidance.
- A ***MaxiMin strategy*** picks the alternative that minimizes the probability of the worst-case outcome. Hence, it was operationalized as taking the maximum of the remaining three attributes in the same direction (positively correlated with each other) but negatively correlated with the attributes used in the MaxiMax strategy. The attributes were fuel efficiency, humanitarian aid, and extra supplies. Here, I take the maximum these attributes instead of the minimum, as the attributes in the study were coded such that a higher number always means better.
- A ***random strategy*** was also added to the model to evaluate if participants were picking the route randomly. A random strategy does not require attribute weightings, as a participant is equally likely to pick any route randomly.

- A ***conjunctive strategy*** requires that the selected alternative must exceed some minimum threshold for each attribute. This was operationalized by taking a minimum of all six attributes for each route to get the utility of that route.
- A ***disjunctive strategy*** requires that the selected alternative exceed some minimum threshold for only one attribute. This was operationalized by taking a maximum of all six attributes for each route to get the utility of that route.

Table 29. Strategy weights for all six attributes : A1- Time Efficiency, A2- Fuel Efficiency, A3- Obstacle Avoidance, A4- Additional Supplies, A5- Weather Hazard Avoidance, A6- Humanitarian Aid

Decision Strategy	**Weights Used for Each attribute for Utility Calculations { A1, A2, A3, A4, A5, A6}**
Equal Weights	{ 1/6, 1/6, 1/6, 1/6, 1/6, 1/6 }
Soft Corrective Weights	{ 0, 1/5,1/5, 1/5, 1/5, 1/5 }
Hard Corrective Weights	{ 0, 1/4, 0, 1/4, 1/4, 1/4 }
Lexicographic 1 (Attribute 1)	{ 1, 0, 0, 0, 0, 0 }
Lexicographic 2 (Attribute 2)	{ 0, 1, 0, 0, 0, 0 }
Lexicographic 3 (Attribute 3)	{ 0, 0, 1, 0, 0, 0 }
Lexicographic 4 (Attribute 4)	{ 0, 0, 0, 1, 0, 0 }
Lexicographic 5 (Attribute 5)	{ 0, 0, 0, 0, 1, 0 }
Lexicographic 6 (Attribute 6)	{ 0, 0, 0, 0, 0, 1 }

Table 29. *Continued*

Decision Strategy	**Weights Used for Each attribute for Utility Calculations { A1, A2, A3, A4, A5, A6}**
Maxi Max	Max of Attribute 1, 3, and 5
Maxi Min	Max of Attribute 2, 4, and 6
Conjunctive	Minimum of all six attributes for each route
Disjunctive	Maximum of all six attributes for each route
Random Strategy	Randomly choosing any route. Each route is equally likely to be picked on each trial.

After calculating each route's utility by every decision strategy based on their corresponding attribute values and weights from Table 29, the probability of participants selected route-choice under each decision strategy was calculated using the traditional choice axiom shown in Equation 6 (Luce, 1959). Following this, the natural logs of the probability for each route choice of a participant for every trial were added to derive the log-likelihood for that participant (see Equation 7) under each decision strategy. The G^2 and Bayesian information criteria (BIC) were then calculated in the standard way by using Equations 8 and 9. The lower the BIC values mean the model is a better fit. Therefore, each participant's best-fit decision strategy will be the one with the lowest BIC out of all strategies.

The same steps from Equations 7-9 were repeated to get the best-fit decision strategy for each participant's every trial block by calculating log-likelihood by each trial block (4 blocks) instead of aggregating over all trials. The same steps were also repeated

to get each experimental condition's best-fit decision strategy by calculating log-likelihood by taking the sum over all trials for all participants in that condition.

$$Luce's\ Choice\ Probability_{for\ each\ Trial} = \frac{Utility\ of\ Selected\ Route}{\sum_{i=1}^{7} Utility\ of\ Presented\ Route} \quad (6)$$

$$Log\ Likelihood\ of\ Choice_{for\ each\ participant} = \sum_{i=1}^{Total\ no.of\ trials} \ln(Luce's\ Choice\ Probability_i) \quad (7)$$

$$G^2 = -2 \times Log\ Likelihood\ of\ Choice_{for\ each\ participant} \quad (8)$$

$$BIC_{for\ each\ participant} = G^2 + k \times \ln(Total\ no.of\ Trials) \quad (9)$$

Best fit decision strategy by condition (aggregated over all participants)

The best-fit decision strategy for each condition in the experiments, their corresponding BIC values, and a BIC value for random strategy are shown in Table 30. A BIC difference between two statistical models of 0 to 2 is not considered enough evidence, 2 to 6 is considered positive evidence, 6 to 10 is considered strong evidence, and >10 is considered very strong evidence in favor of a model with lower BIC value (Liao & Fasang, 2021). The BIC difference for best-fit strategies in Table 30 is much >10 for all conditions compared to their corresponding random strategy BIC values.

Table 30. Best fit strategy and BIC values for each condition

Model Misspecification	Recommender	Best Fit Decision Strategy	BIC- Best Fit Decision Strategy	BIC- Random Strategy
Accurate	No	Equal Weights	4533.56	4677.27
	Yes	Conjunctive	4497.99	4657.81
Soft	No	Soft Corrective Weights	4481.93	4673.38
	Yes	Soft Corrective Weights	4486.47	4673.38
	Explanation	Soft Corrective Weights	4682.71	4829.09
Hard	No	Equal Weights	4438.68	4669.48
	Yes	Hard Corrective Weights	4622.14	4657.81
	Explanation	Hard Corrective Weights	4477.03	4813.52

I predicted that participants in hard misspecified conditions would be better fit by EQ weights compared to soft and accurate for both no and yes recommender conditions. This prediction was counter-intuitive and based on the reasoning that participants in the hard condition would have to put the most effort into finding the best route due to misspecifications. On the other hand, accurate condition participants can still have an optimal performance by using less than optimal strategies due to the near-perfect accuracy of the MDSS. The support for this evidence is only present for no recommender hard condition which has EQ weights as the best-fit strategy and lowest BIC out of three

misspecification levels. The participants in the accurate recommender condition were best-fit by the conjunctive strategy. The participants in all soft conditions were fit best by the soft corrective weights. The participants in recommender and explanation hard conditions were also best fit by their respective hard corrective weights. This additional effort by participants in misspecified conditions of the experiment using compensatory corrective weights maybe explains why there were not many performance differences between different model misspecification levels. As proposed in the model-blindness framework, participants' strategies might compensate for model limitations to some extent, thus model-limited performance degradations in hard and soft conditions can be mitigated. I expected participants to fit better by their respective corrective weights for explanation conditions. There was evidence for this prediction for both hard and soft explanation conditions. Hence, participants' best-fit decision strategy does reflect their efforts to incorporate the explanation message.

Best-fit decision strategy by participants

Figure 53 shows the percentage of participants fit different decision strategies included in the analysis by all conditions in both experiments. It is worth mentioning here that equal weights, conjunctive, soft corrective weights, and hard corrective weights are all compensatory strategies taking into account most attributes in decision-making; hence, they are also statistically very similar and competing strategies in estimation. In all conditions, the maximax, maximin, disjunctive, random strategy, and most lexical strategies fit the least number of participants. Figure 54 shows the same data as Figure 53 by grouping strategies based on their characteristics. Participants fit by equal weights, soft corrective, hard corrective, MaxiMin, MaxiMax, and conjunctive strategies were grouped

as participants following compensatory strategies. Participants fit by disjunctive, and any of the lexicographic strategies were grouped as participants following non-compensatory strategies.

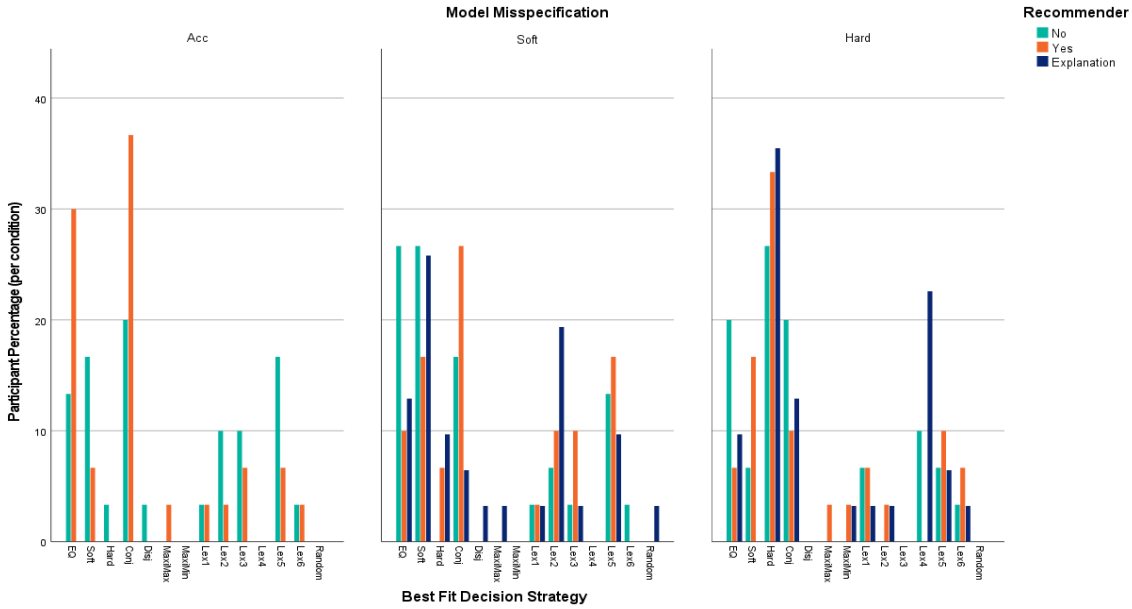

Figure 53. Percentage of participants fit by different decision strategies by conditions

More participants in the accurate condition with recommender used compensatory strategies compared to non-compensatory or random in comparison to control condition participants. Hence, presenting highly accurate recommended routes helped participants to use more optimal strategies to improve performance. On the other hand, for participants in soft conditions, there was a slight increase in the proportion of participants using non-compensatory strategies in recommender and explanation conditions compared to the no-recommender condition. For hard conditions, explanation availability made more number of people follow less than optimal non-compensatory strategies compared to both control

conditions (yes and no recommender). This might indicate a negative impact of presenting an explanation which can lead to overconfidence and the use of less-optimal strategies by some users. For no recommender conditions, the accurate condition had the lowest percentage of participants following compensatory strategies compared to soft and hard. This trend was reversed for recommender conditions as the accurate condition had the highest percentage of participants following compensatory strategies, then hard, and the soft had the lowest percentage of participants using compensatory strategies. Explanation conditions had an almost identical number of people in both soft and hard conditions using compensatory strategies.

These observations from decision strategies are interesting, given that there were no significant differences in participants' performance levels in most experimental conditions. Hence, participants in this dissertation study overcame some level of model-limited performance by using better decision strategies (reducing strategy-limited) performance.

For Experiment 1, I predicted that participants in hard misspecified conditions would better fit by equal weights compared to soft and accurate for both no and yes recommender conditions. No consistent evidence for this prediction was found as there are a lot of individual differences in participants' best-fit strategies. I also predicted that there would be greater variability in decision strategies among participants in the hard condition compared to soft and accurate conditions. Participants had to overcome the misspecification and find the best strategy to improve performance. However, no consistent evidence for this prediction can be seen in Figure 53.

For Experiment 2, I predicted that participant decision strategies for explanation conditions would mimic their respective corrective weights more closely compared to their corresponding control conditions from Experiment 1. This was because the explanation message directly indicated participants to underweight attributes by explicitly spilling information about how they were overweighted in the MDSS algorithm. Both hard and soft explanation conditions have more participants fitting by either actual corrective weights (soft/hard) or lexical strategies compared to their controls from Experiment 1. Thus providing an indication of the positive impact of providing an explanation even though much evidence about performance improvements was not found.

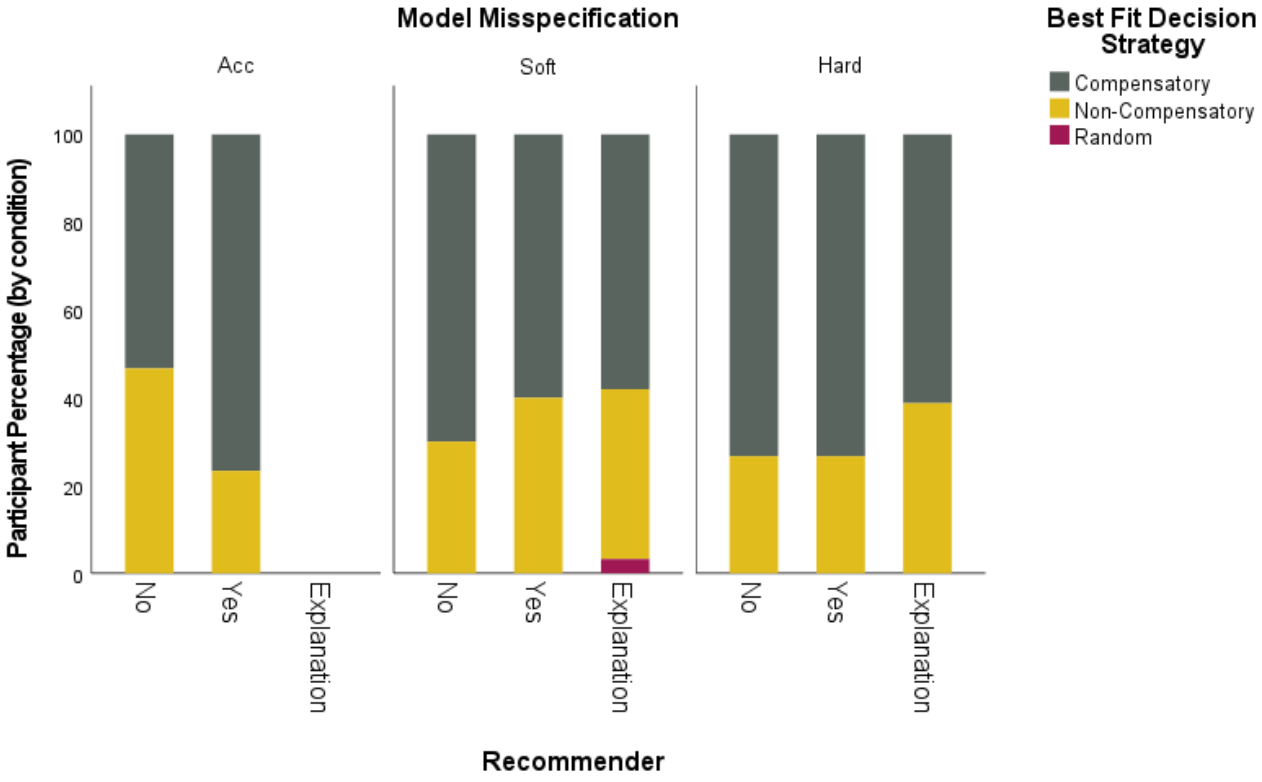

Figure 54. Percentage of participants fit by either compensatory or non-compensatory strategies or a random strategy by conditions

Best fit strategy by participant and trial block

I predicted that decision strategies would better fit the actual model weights (respective WADDs- EQ, soft, and hard) as a function of the trial block progression due to learning. However, no clear evidence can be found for this prediction in Figure 55 and Figure 56. It is also evident from the learning results presented before. There was no significant performance difference found with trial block progression.

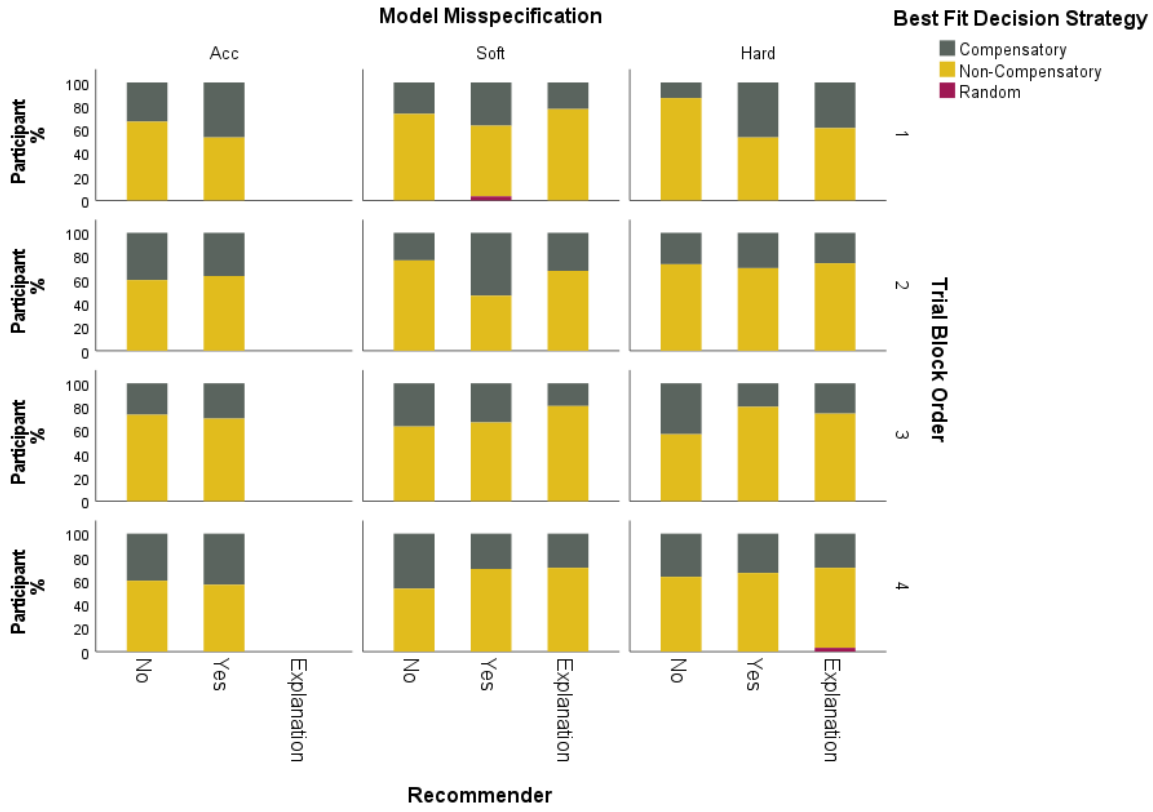

Figure 55. Percentage of participants fit by either compensatory or non-compensatory strategies or a random strategy in each trial block by conditions

In all trial blocks in Figure 56, the majority of participants are now fit by non-compensatory strategies. However, it is important to note here that strategy estimation via

block is more prone to model estimation errors and should be trusted less than estimation via aggregating all participants or conditions due to very few data points available in each block. This might be why the results are completely reversed compared to strategy fits by participants and by conditions shown before.

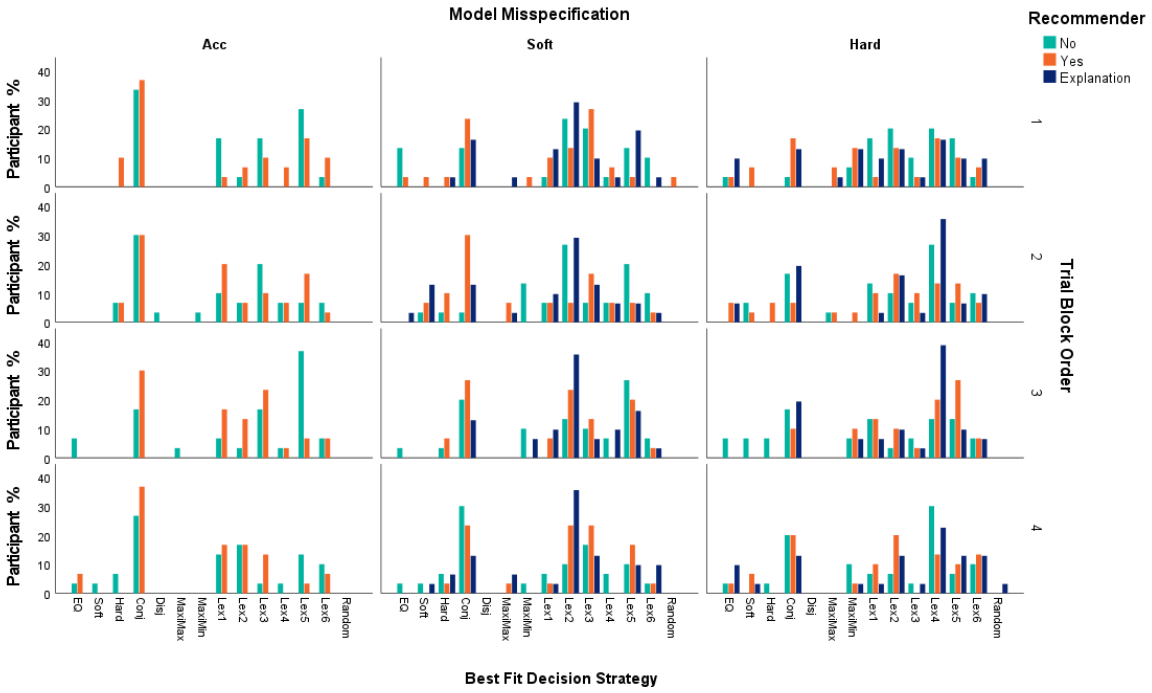

Figure 56. Percentage of participants fit by different decision strategies in each trial block by experimental condition

3.5 Participant's Actual Attribute Weights in the Task as Decision Strategy

After estimating the best-fit decision strategies for participants from a list of pre-determined strategies or attribute-weighting schemes, I also freely estimated the actual weights each participant gave to each attribute. The attribute weights were determined by using data for all seven routes presented on all forty trials to the participants. A participant's selected route was given a score of 1, and the remaining 6 routes on each trial were given a score of 0 to get the route-selection DV. A linear regression model with route selection

as DV and values of six attributes as predictors was run without an intercept for each participant separately. The obtained standardized regression coefficients (β) for each attribute were equivalent to the weight given to that attribute by the participant. The β-weights should be almost equivalent for all attributes for a participant following an equal weights strategy. Figure 57 shows the mean β-weights for participants by condition.

On average, participants in the soft explanation condition negatively weighted (underweighted) "time efficiency," as instructed by the explanation message. A similar trend was also observed for participants in the hard explanation condition as they, on average, negatively weighted (underweighted) "time efficiency" and gave a negligible weight to the "obstacle avoidance" attribute, as instructed by the explanation message. Hence, the explanation did help participants shift their attribute weights more toward respective corrective weights.

Participants also demonstrated a general bias towards some attributes, leading them to underweight extra supplies and humanitarian aid attributes for accurate and soft conditions, and humanitarian aid attribute for the hard condition. This bias led to participants' deviation from the equal weights strategy in Experiment 1 conditions. This probably indicates the sensitivity of participants' to attribute label names, leading them to ignore attributes that sound less important based on the task context of delivering supplies.

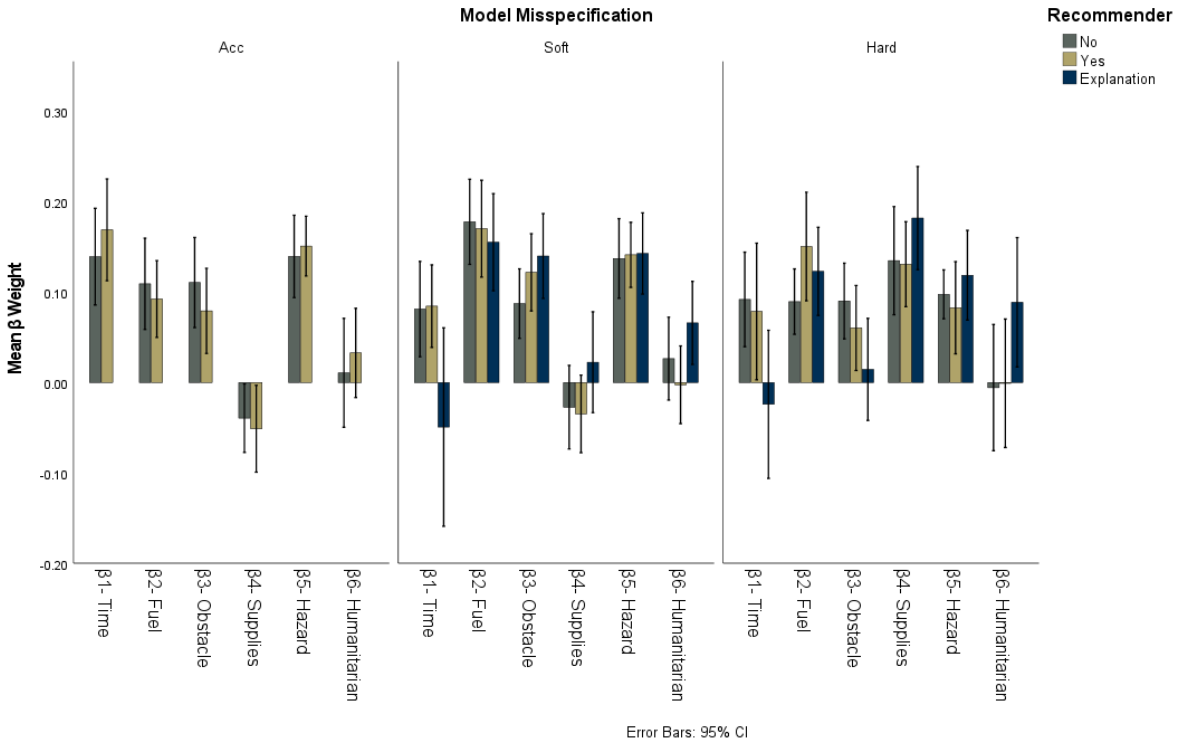

Figure 57. Mean regression coefficients of participants' route choices regressed on attributes by condition

3.6 Feedback and Experience Questionnaire Analysis

Participants were asked to rate their experience with playing video games and how often they follow recommender suggestions (See APPENDIX B and C for the questionnaire). The experience level was correlated with participants' performance in the task. Participants were also asked to respond to free-text feedback and experience questions about the route-recommender system, which were analyzed using thematic analysis using multiple raters.

3.6.1 Rank order correlation between choice rank (performance) and experience level

Participants rated on a 5-point Likert scale (Very often, often, occasionally, rarely, never) how often they play video games and follow recommender system

recommendations. A rank order correlation shows a small significant negative correlation (τ= -0.114, p=0.024) between participants' experience playing video games and their mean route-choice rank. That means performance gets better (rank decreases) with the increase in video game experience. A rank order correlation also shows a small significant positive correlation (τ=0.143, p=0.012) between participants' experience playing video games and their recommender use. The correlation between recommender use and performance (τ=-0.003, p=0.95) was not significant.

3.6.2 Content Analysis Results

Content analysis (thematic analysis) was conducted for the free-text open-ended responses of participants to three questions in Experiment 1 and five questions in Experiment 2 (see APPENDIX C for the post-study questionnaire). The questions and corresponding coding categories are shown in Table 31. Table 32 shows an example of participant responses to AI limitations (Question 1) and consequent performance impact (Question 2) questions for all conditions in both experiments. Two researchers worked together to develop the themes for each question by following both the top-down and bottom-up approaches. The top-down approach focused on study design and question context to develop themes relevant to the question asked. The bottom-up approach focused on looking through a few participant responses to refine the previously identified themes and add more themes. Each identified theme served as a coding category. The researchers tried to create categories that are as mutually exclusive as possible so that raters can classify a participant's response into only one category with the least conflict.

Four undergraduate researchers acted as four raters (R1, R2, R3, and R4) to classify responses into categories. Four raters coded all questions in random pairs of two-raters coding 50% of answers for each question. Raters were blind to the knowledge about which experimental condition each response belonged. Cohen's kappa (κ) was calculated to measure inter-rater reliability between two raters. Two kappas were calculated for two pairs of raters, who coded 50-50 data for each question (Table 33). The κ of 0.4-0.74 is usually considered moderate to good agreement between raters (Altman, 1990; Landis & Koch, 1977). Based on this range, most of the rater pairs had moderate to good agreement, except for the first half of Question 4 and the second half of Question 5, which had an unacceptable level of agreement ($\kappa<0.4$). For these two response sets, a third rater, R3 or R4, was assigned to code both questions again. Their kappas with the previous raters are also shown in Table 33. The coded data from the third rater was selected for both questions as the third raters seemed to have a moderate agreement with the other two raters. The rater R3 usually had the highest agreement with any other rater for all other questions. Hence, whenever R3-coded responses were available, data were selected from R3. Otherwise, coded data was randomly picked from the other three raters.

Table 31. Themes/data coding categories identified for each question

#	Questions	Coding Categories
1	Please describe any limitations/issues you noticed with the AI-based route recommender system.	No errors or issues with AI Inconsistent/unreliable / Inaccurate system/Lack of transparency Biased or unknown weights Confusing/incorrect visual display graphics More information needed No response from participant Other
2	How did the limitations you listed above (if any) affect your performance in selecting the best route in the experiment today?	No impact on performance Ignored AI/Didn't follow AI/ Used own judgment Affected trust in system/ confidence in decision Described decision strategies used (participant's strategy) No response from participant Other
3	What additional information would you prefer to receive about/from the AI-based route recommender system?	No additional information needed Weights/Sums/Average Better feedback on performance Explanation about AI No response from participant Other
4	Please describe any limitations/issues you noticed with the "warning message" you received about the error in the AI-based route recommender system provided to you.	No issues with warning message Misleading Repetitive Inadequate/confusing/vague No response from participant Other
5	What additional information would you prefer to receive about the AI-based route recommender system via the "warning message"?	No additional information needed More explanation about bias Direction/magnitude of bias Weights No response from participant Other

Table 32. Example participant responses to limitation and performance impact question

Model Misspecification	Recommender	Participant comments [Limitations; Performance Impact]
Accurate	No	*"The system needs more evidence to make it more trust worthy. Also, I feel like in each round of the game the priorities were different, I could not find a pattern to help me choose the best outcome."*
		"It made it hard to decide which route was the best because I had no previous knowledge of which qualities of the route plan are more important."
	Yes	*"i noticed it was hard to figure out the AI's model for picking the best path, even after analyzing many scenarios. it was hard to figure out what the recommender valued"*
		"my performance was affected because it was hard to know what categories the route recommender valued over others."
Soft	No	*"There were no issues, I was not sure whether some of the categories had more weight than the other categories because if that was the case, I would have made my decisions differently."*
		"I tried figuring out which set of numbers had the higher average, some sets confused me because there were multiple zeros so I tried to go with the option that had the most consistent and somewhat high set of numbers. Sometimes it was the best option and other times it was not."
	Yes	*"There were times in which the AI system would recommend routes with more 0 attributes than "high" (80-100) attributes, which made me a little wary. There were also times where I selected a recommend route and it would turn out to be the sixth best route, and other times where a non-recommended route would be second or first. At times, it felt like the recommended routes were randomly selected."*

Table 32. *Continued.*

Model Misspecification	Recommender	Participant comments [Limitations; Performance Impact]
Soft	Yes	*"It affected my performance as I began to solely rely on the numbers, only using the recommender if I was truly stuck."*
	Explanation	*"I didn't know how much more importance was placed on time-efficiency in comparison to the other variables."*
		"It affected my performance in choosing as I didn't fully know when to decide between a low time efficiency value route and route with high values for other variables."
Hard	No	*"I feel like at times it maximized for high averages (no zeros), and other times it looked for the highest raw summation"*
		"I'd just generally try to roughly add stuff up, early on I was discounting routes which had 0s in certain areas (like hazardous avoidance), but later I just looked through everything and counted up which route had the highest +90 answers/highest raw estimate"
	Yes	*"Personally I tried to avoid options with 0s because I was aiming for a balance - I noticed that at times the AI did not take this route of reasoning."*
		"Sometimes I would disagree with the AI's routes and disregard them after seeing so many 0s - it essentially hindered the AI's reliability at times as well."
	Explanation	*"I was unsure how much bias was impacting the rating values for the two affected fields."*
		"Often I didn't select the most optimal routes because I believe I was unable to properly gauge how much bias was impacting the AI outputs values. Typically, the AI routes would produce the highest scores, but it was hard to gauge how that translated to an actual score without bias"

Table 33. Inter-rater reliability between raters for all questions for 50% data

Question	Raters	Number of responses	Cohen's Kappa (κ)	% agreement	Data selected
1	R1, R2	108	.477	58%	R1
	R3, R4	108	.492	60%	R3
2	R1, R4	108	.494	60%	R4
	R2, R3	108	.537	66%	R3
3	R2, R4	108	.528	64%	R2
	R1, R3	108	.671	76%	R3
	R1, R4		.378	52%	
4	R1, R3	31	.597	71%	R3
	R4, R3		.534	64%	
	R2, R3	31	.746	71%	R3
5	R3, R4	31	.537	61%	R3
	R1, R2		.315	42%	
	R1, R4	31	.481	58 %	R4
	R2, R4		.466	58 %	

Question 1: Limitations/issues with the route recommender system

Participants' responses to Question 1 indicated they found two major limitations of the recommender system irrespective of the experimental condition (Figure 58). Firstly, participants felt that the system was inconsistent, unreliable, inaccurate, and was not transparent. Second, participants also felt that the system was using biased or unknown attribute weights in weighing the presented route attributes. Figure 58 shows that the proportion of participants who felt attribute weights were biased in explanation conditions was much higher than in corresponding control conditions. This might be attributed to the explicit and direct explanation message they received. Interestingly, the proportion of participants who requested more information about the recommender system are the ones

who received recommended routes compared to participants in no recommender control conditions.

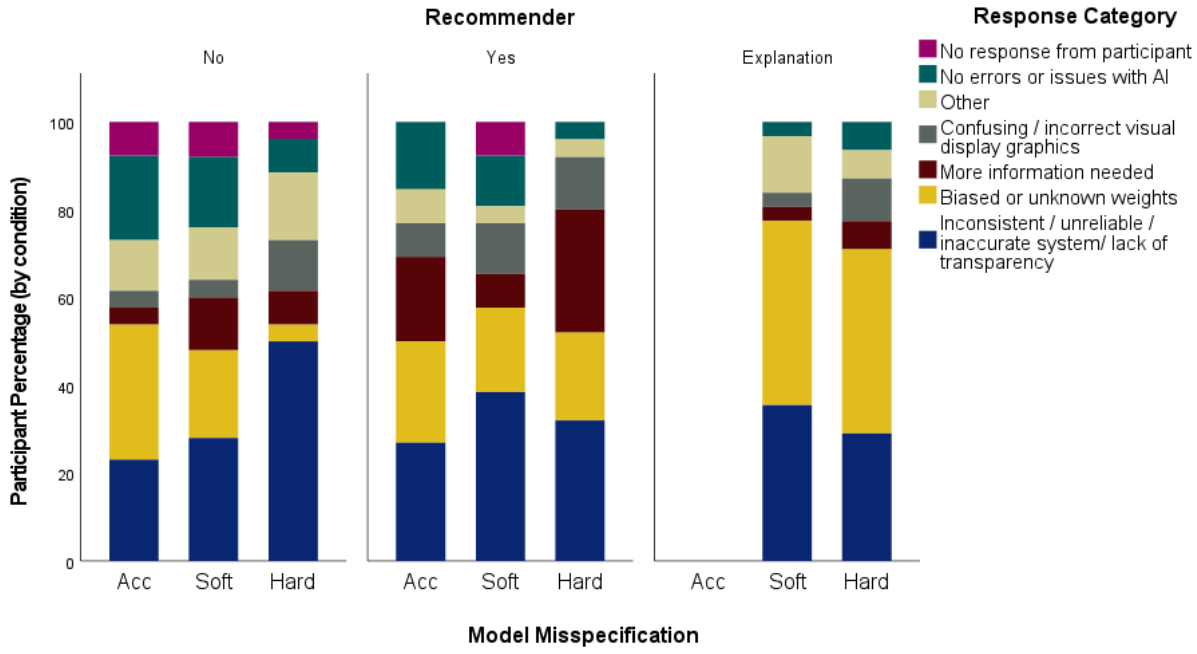

Figure 58. Limitations/issues identified with the route recommender system via open-ended participant responses

Question 2: Impact of limitations on participant performance

Participants' responses to Question 2 asked them to talk about the performance impact of the limitations they identified in the system. Most Participants talked about how the system negatively impacted their performance, trust in the system, or confidence in their decision. Many participants also described what weighting schemes (decision strategies) they used to evaluate routes and associated attribute weights to compensate for those system limitations. Figure 59 shows the proportion of participants in each condition and what category their response belonged to. More participants in the hard conditions (at all levels of recommenders) said that they ignored or didn't follow the recommender system and used their judgment compared to the soft and accurate conditions. This is

consistent with the prediction that it was easier for participants in the hard conditions to detect the issues with the system than in the soft conditions.

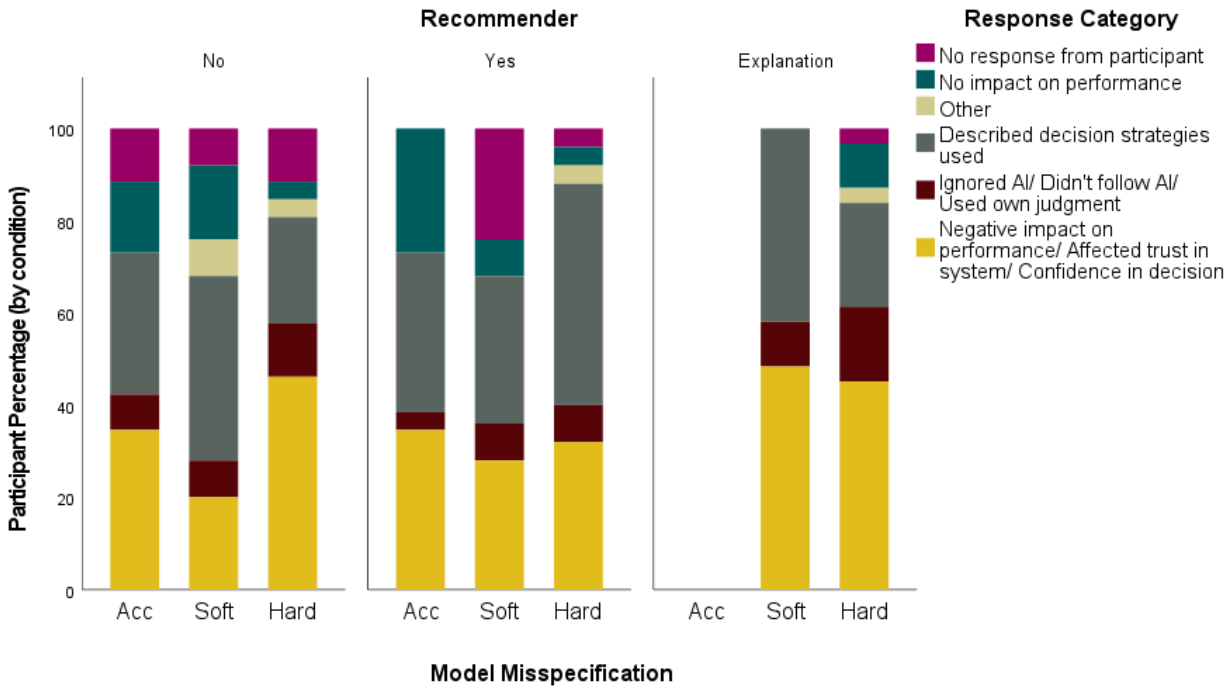

Figure 59. Performance impact of limitations with the route recommender system via open-ended participant responses

Question 3: Additional information needed about the route recommender system

Most participants indicated they needed more details or an explanation of the AI algorithm. The majority of them also indicated a need for exact weights for attributes or sum or averages for each route across the six attributes for easy comparison and decision choice. Figure 60 shows the proportion of participants in each condition and what category their response belonged to. More participants in the soft condition indicated a need for better feedback on performance compared to the accurate and hard conditions for both yes and no recommender levels.

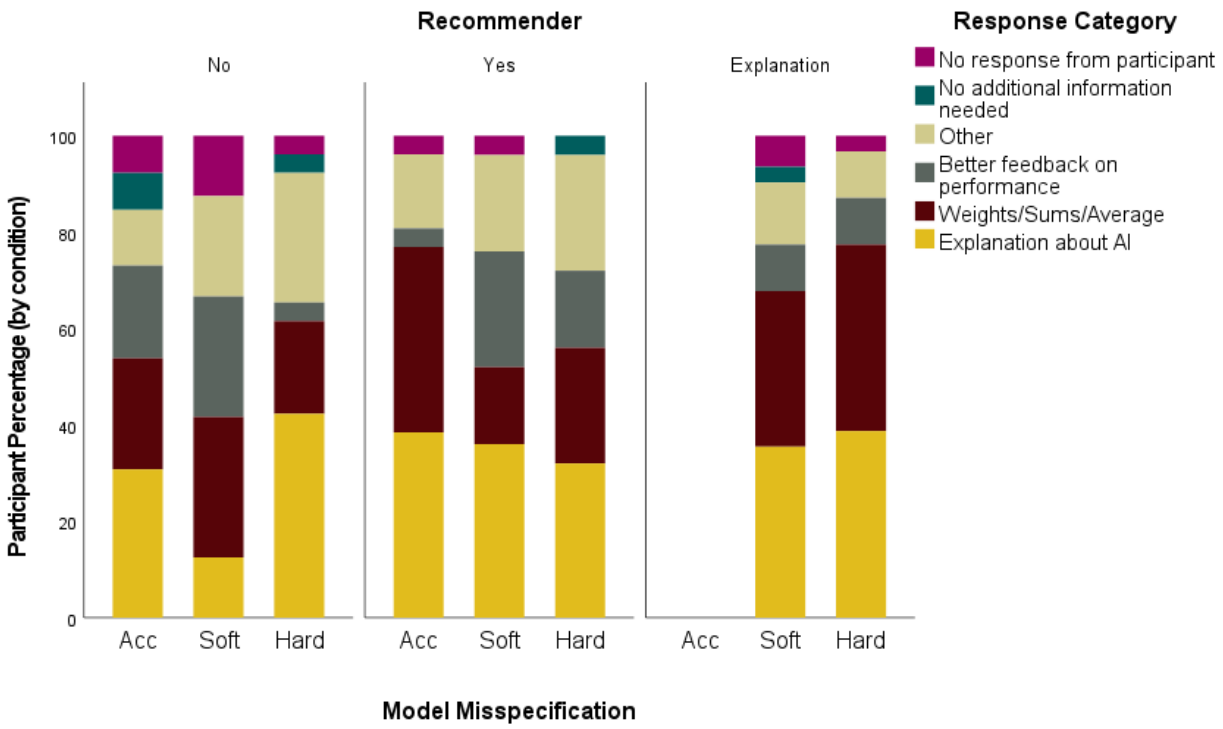

Figure 60. Additional information requested by the participants about the route recommender system via open-ended responses

Question 4: Limitations/issues with the explanation message and Question 5: Additional information needed in the explanation message

Only participants in the explanation conditions responded to Questions 4 and 5. Figure 61 and Figure 62 show the proportion of participants and what category their responses belonged to for both questions. Even though we believed that the explanation message was very straightforward and provided more information than conventional XAI interventions provide, most participants, especially for the hard condition, found that the explanation message was inadequate, vague, and confusing (Figure 61**)**. Many participants also asked for more information about the bias or exact attribute weights, magnitude, or direction of the bias in both soft and hard conditions **(**Figure 62**).**

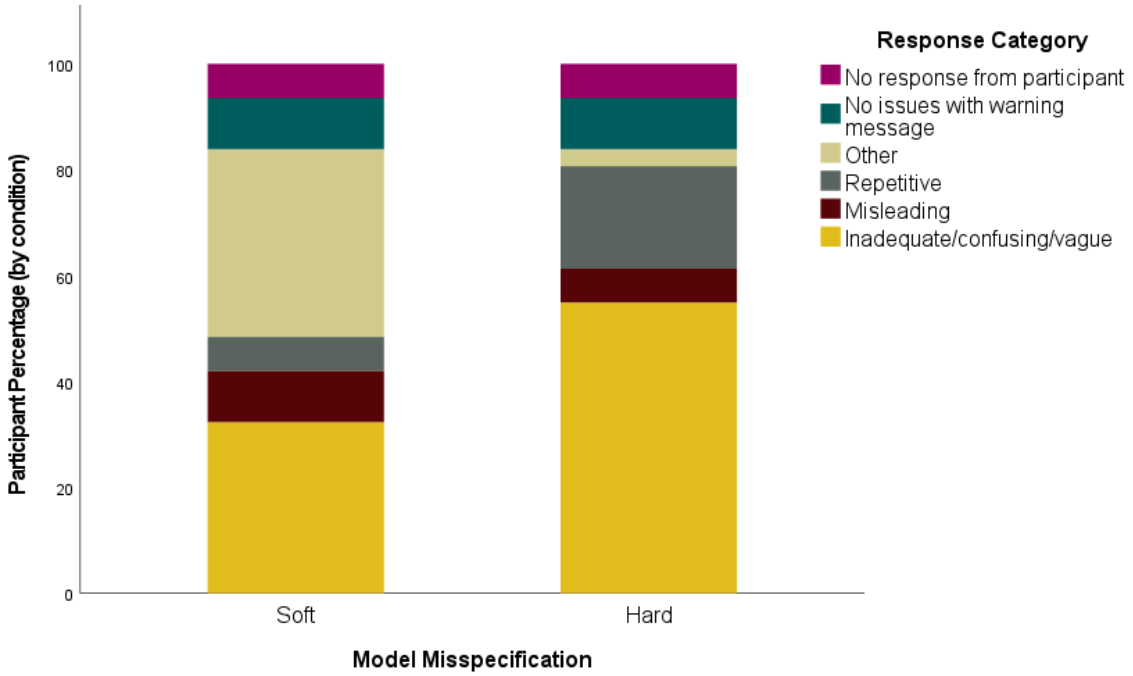

Figure 61. Limitations/issues identified with the explanation message via open-ended participant responses

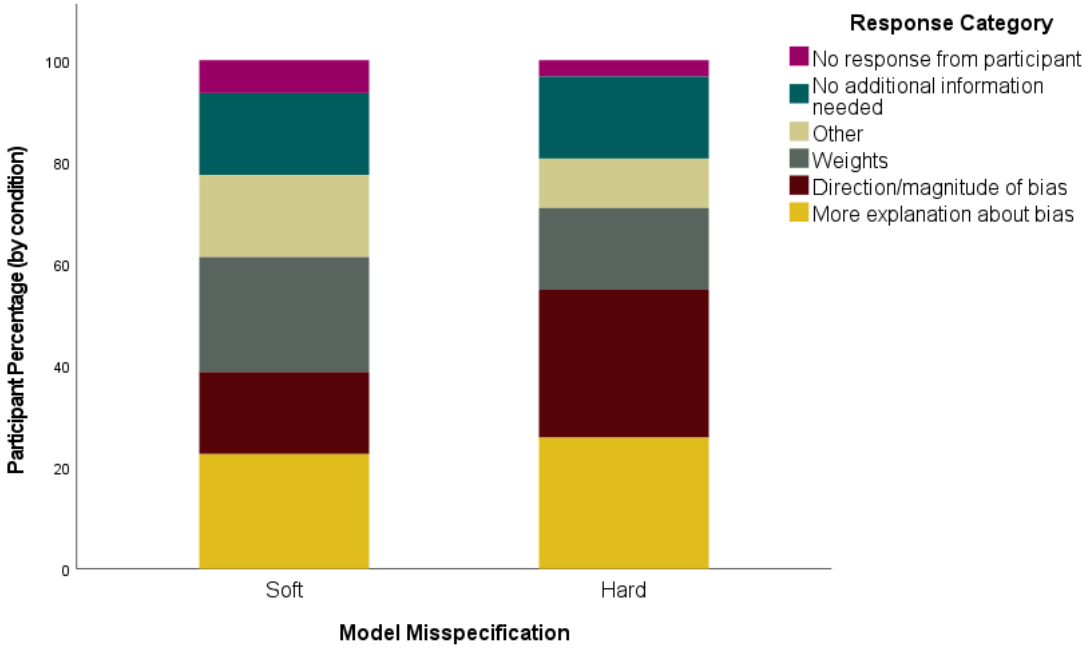

Figure 62. Additional information requested by the participants in the explanation message via open-ended responses

Overall, the content/themes analysis provides some interesting insights into how participants of these dissertation experiments understood the misspecifications in the system and errors in the recommender and adapted their decision-making process. It also provided insights into what worked or didn't work for the route recommender system and explanation messages. This qualitative data helps explains some of the quantitative results presented before. These conclusions will be discussed more in the general discussion section.

CHAPTER 4. DISCUSSION

The primary goal of this dissertation was to formally investigate the role of model blindness imposed by model-based decision support systems implemented in dynamic, uncertain, multi-objective, and high-consequence decision-making environments. The dissertation also examined a mitigation technique for misspecified models to calibrate users' decision-making under model blindness. The aim was to empirically evaluate our proposed model blindness confluence model (Parmar et al., 2021), which was supported via simulations presented earlier in Chapter 2. The confluence model comprised three components— the MDSS, the task's decision context, and the user's decision strategy. Each component can potentially lead to performance degradation.

4.1 Results Summary and Discussion

Experiment 1 of this dissertation took the first step in elucidating how model blindness can manifest in a user's decision-making process by carefully manipulating model misspecification and recommender availability and evaluating various performance measures, trust, confidence, and decision strategies. The route recommender system also provided a unique capability to participants, allowing them to deviate from the recommended set by presenting additional alternatives. The results only provide partial evidence for participants' performance degradation due to model blindness. This experiment does not provide strong support for many of the hypothesized effects related to model misspecification levels. The performance improvement (rank and local utility loss) was observed for participants only when recommended routes were presented to them compared to participants without any recommender. Global utility loss was the only

measure that differed across model misspecification levels. Although this global level performance difference is unrelated to local measures like trust, confidence, etc., this can be a manifestation of trust miscalibration which will be discussed later. Participants' Brier scores were also found to be only calibrated when the system was highly accurate compared to misspecified systems. There was no difference between soft and hard conditions' Brier scores, even though the recommender system differed in accuracy between those conditions.

Interestingly, the predicted interaction effect for model misspecification and recommender was only observed for the reliance measure. Participants were likely to rely on recommender (pick routes from the recommended set) when it was present compared to when it was not for the accurate and soft conditions. However, for the participants in the hard conditions, there was no reliance difference between those with and without the recommender. Hence, participants with and without recommender were equally likely to pick from those three routes whether it is recommended or not. This can be interpreted as no effect of recommender for participants in the hard condition. This finding is interesting as the prediction was that participants in the hard condition with the recommender would find it easier to detect the recommender's poor quality and avoid over-reliance. In contrast, the prediction for the participants in the soft condition with recommender was that slight misspecification could be more detrimental for users as it might lead to over-reliance due to the recommender being correct sometimes.

Experiment 2 implemented a unique way of presenting XAI-type explanations to reduce the impact of model blindness imposed on users in the misspecified conditions from Experiment 1. The results of Experiment 2 are similar in trend to the Experiment 1 results.

Experiment 2 only provided partial evidence for the success of presenting explanations to create model blindness awareness among users. The results did show awareness among participants through decision strategies, but no reliable performance differences were observed between conditions with or without explanations. The Brier score calibration followed the same patterns as Experiment 1, with no calibration differences between soft and hard misspecification levels. The participants' reliance on the recommended set in the explanation conditions does not differ from those without the recommender. However, participants' reliance on the recommended set in the explanation conditions was significantly lower than participants in the recommender conditions without the explanation. Hence, providing explicit explanations does reduce reliance on a misspecified system even though no performance changes were detected between the levels of misspecification. This provides evidence to support that presenting natural language explanations can help calibrate decision makers to the capabilities and limitations of an MDSS, although it didn't improve performance.

Explicit explanations also played a role in participants' trust score calibration, as trust decreased compared to the conditions without a recommender. Surprisingly, participants' trust in the explanation conditions significantly reduced from no recommender to explanation conditions for the scale's trustworthiness, technical competence, and personal attachment facets. The understandability and reliability facets were not different between any conditions of both experiments. Hence, trust turns out to be a concerning and counter-intuitive measure, as providing an explicit statement about bias was the only form of 'feedback' that significantly reduced participant trust in the system. However, this is not the first time we have seen these issues with trust measures;

in one of our previous studies (Parmar & Thomas, 2020), people were highly sensitive to DSS's accuracy and showed significant performance and Brier score differences with accuracy changes. However, trust scores still didn't change between conditions in that study, and they remained at neutral levels with a 5-point Likert scale. The follow-up exploratory analysis in that study also showed that trust was only found to be calibrated to operators' objective performance when the tool was highly accurate. Learning from those findings, we chose a different and more appropriate measure of trust for the given task context and also used a 4-point forced Likert scale without the neutral option of 'Neither Agree nor Disagree". However, trust still didn't change even though participants' global utility loss (a measure of absolute tool accuracy) significantly differed between conditions in both experiments.

Trust and reliance are considered to be positively correlated measures in trust literature, as reliance (or use) of any automation depends on its trustworthiness (Lee & See, 2004; Muir & Moray, 1996). In this study, there were significant changes in reliance between conditions for both experiments; however, the trust didn't vary as expected. To further test how trust and reliance are related in our task, a Pearson correlation was calculated between mean reliance and omnibus trust score across participants. The two were not significantly correlated, $r=0.036$, $p=0.578>0.05$. There was also no significant correlation between trust and performance, $r= -0.094$, $p=.149>0.05$. A study by Dzindolet et al. (2003) showed the results opposite to what we found. In their study, authors found that providing an explanation about situations when an automated aid presenting target present vs. absent (binary choice) decision support might fail (can give false alarms) and telling participants that the aid is not perfect led to improved reliance and trust compared

to no explanation conditions. Interestingly, they also found that participants considered the trustworthiness of a superior aid higher than an inferior aid, but they were found to rely on both equally likely. I think, in our case, reliance changes and reduces with explanations because we have presented participants with additional alternatives to choose from when they feel the system is making poor recommendations. Reliance changes in this dissertation study might also be attributed to an explanation message explicitly telling participants that MDSS is always misspecified, instead of telling them, like Dzindolet et al. (2003), that aid might be wrong on some trials. Based on these results, it seems as though the construct validity of trust in automation measures is of concern (Stuck et al., 2022).

No reliable and consistent learning effects across trial order or block order were observed in either experiment. The lack of an observed learning effect might be attributed to differences in the difficulty level across the different trial scenarios. The trials (scenarios) show significantly different performance across misspecification and recommender levels. Even though performance, trust, confidence, reliance, and learning provided partial support for the hypothesized effects, the decision strategy analyses provided useful insights into participants' decision-making processes that could explain some of those results. The decision-strategy model-fitting analysis aggregated by condition showed the participants' route choices were best-fit by the equal weights, or corrective weights for misspecified conditions, and conjunctive and equal weights for accurate conditions. Hence these results indicate that decision-makers adapt their strategies when there is bias or misspecification in MDSS that can be detected from available feedback. The participants adjusted their decision strategy to either match true world feedback (non-explanation conditions) or overcome bias (explanation conditions). The same trend was

also observed from participants' actual β-weights obtained from their route choice and corresponding attribute values presented in the task using linear regression. On average, β-weights for participants in the explanation condition showed that they could incorporate explanation messages into their decision-making to approximately move towards respective hard and soft corrective weights. Participants in all non-explanation conditions showed sensitivity towards attribute label names in their average β-weights, causing them to underweight some attributes that might sound less important for the decision task of delivering the shipment. Participants in those conditions were still weighing other attributes almost equally.

The decision-strategy model-fitting analysis by participants showed that more participants used non-compensatory strategies in explanation conditions compared to no-explanation control conditions. This might explain why there weren't any performance improvements with explanations, as some participants found our straightforward explanation message challenging to incorporate into decision-making. This was evident from their post-study questionnaire response, in which most participants found the explanation message misleading, confusing, and vague and requested to receive more information. Hence, these results indicate a need for XAI researchers to consider the role of decision strategies more carefully when implementing any XAI capability in systems. The accurate condition with recommender had the most participants using compensatory strategies compared to the hard condition with recommender, and the soft condition with recommender had the least number of participants using compensatory strategies. The accurate condition with recommender afforded participants the highest chance for success with additional effort, followed by the hard misspecification, most likely because the hard

misspecification was easier to detect compared to the soft misspecification. The trend was reversed for the accurate condition without recommender (controls) as it had the lowest percentage of participants using compensatory strategies compared to hard and soft (controls).

4.2 Conclusion

Although, the results of this dissertation were not conclusive in terms of performance changes. It revealed some important characteristics about how users are prone to model blindness and how their trust in the system only changes when they are explicitly told that the system's algorithm is biased towards specific attributes. Also, their confidence is only calibrated to accuracy when the system is highly accurate. The decision-strategy analyses and simulation results in this dissertation provided further evidence for the important role of the decision-strategy component of Parmar et al.'s (2021) proposed model blindness confluence model. Participants also frequently described their detection of an error in AI and how they changed their decision strategies to improve performance on the task in the post-study questionnaire, indicating participants' ability to adapt to model misspecification levels. Overall, the experiments provide little support for model-limited performance due to model blindness imposed by misspecified systems. The behavior captured in Experiments 1 and 2 showed minimal sensitivity to the different misspecified statistical environments participants operated within. There was evidence of the impact of recommender presence and reliance. The participants' strategies showed that they could understand model limitations and adjust their strategies accordingly to account for misspecifications in the model. Hence, not enough evidence was found for how performance

will be strategy and model-limited under an unfriendly context-limited environment with severe tradeoffs operating.

Finally, I think the motivation, caliber, experience level of participants (experience with recommenders and video games), and background (the majority were CS majors) also played a significant role in good performance levels in all conditions of both experiments. This is also evident from the quality of detailed responses in the comments they added to the post-study questionnaire, where most participants asked for more information about the attribute-weighing scheme of the MDSS algorithm. The results also indicate that people can adapt their behavior to misspecifications to achieve a good performance level even though they might not have metacognitive awareness about the misspecifications. Even though calibration (Brier score) was tied to performance but ideally, both raw confidence score and trust were not. The lack of differences between those measures indicates a lack of metacognitive awareness about misspecifications. Hence, participants can still suffer from model blindness but achieve good-enough performance levels. Finally, this dissertation's results also help understand the challenges with measuring and evaluating trust in a model blindness situation, as having high or low trust might not reflect anything about reliance, performance, or confidence.

4.3 Limitations and Confounds

Since this study was unique and among the first ones to investigate model blindness confluence model in an experimental paradigm, most of the limitations in design can be realized in hindsight. In hindsight, one limitation I see is a significant difference in the difficulty level of some scenarios, which was reflected in participant performance. Hence,

detecting any performance changes or learning effects between conditions was difficult. I could have controlled for that in the experiment design instead of randomly generating scenarios following the required statistical structure of the task. However, in real-world automation, the difficulty levels do change significantly based on task conditions. Usually, automation fails when an abnormally tricky situation occurs, affecting trust, reliance, and confidence in trust in automation literature (Dzindolet et al., 2003; Dzindolet et al., 2001; Johnson et al., 2004; Muir, 1994; Muir & Moray, 1996). Another extraneous variable that could have been handled more carefully was labeling attributes. Because the decision strategy analyses showed participants were underweighting attributes like "extra supplies" and "humanitarian aid" in some conditions as it sounded less important than other attribute labels like "time", "fuel", "hazards" and "obstacles". I think calling the shipment of healthcare supplies in instructions might have led to this bias, even though the same instructions told participants that all attributes are equally important.

CHAPTER 5. IMPLICATIONS AND FUTURE RESEARCH

The results of this dissertation provided partial empirical support for the confluence model of operator performance proposed by Parmar et al. (2021). However, the results did provide insights into trust and confidence-related barriers in MDSS implementations. It also showed the important role decision strategies could play in the success or failure of any XAI manipulations. Although performance changes were not detected, this might be attributed more to the experiment design and range restriction in our participants. It does not indicate that model misspecification didn't affect decision-making, as other measures did provide partial support for some of the hypothesized effects.

5.1 Implications

This dissertation research has theoretical and practical contributions as it combines research from decision-making, human factors, and XAI literature to propose a model blindness confluence framework and empirically test it. Through decision-strategy fits, the dissertation results demonstrated that there might be negative consequences to making users aware of model blindness through XAI explanations, as some participants were found leaning towards non-compensatory strategy fits in those conditions. Hence, it presents researchers with decision-strategy fitting as a tool to evaluate any XAI intervention before system-level implementations. The reliance and trust results also help establish a need for providing additional information and alternatives to users instead of providing a forced recommendation for more thoughtful decision-making. It will also help avoid over-reliance on automated systems.

The results also support the argument raised in Parmar et al. (2021) that model blindness should be considered before adopting an MDSS technology in a high-consequence and multi-objective decision-making environment; as we can see, there is a range of variables that needs to be considered for appropriate adoption of any MDSS. The results also show an advantage of, at minimum, making human operators aware of issues related to model blindness so they can move towards more corrective decision strategies. Some of the formal evaluations of model blindness presented in this dissertation, along with subsequent follow-up experiments in the future, can become a part of the design philosophy used by model and system developers. Model blindness evaluation, awareness, and consequent mitigation will eventually provide ways to build Responsible, Equitable, Traceable, Governable, and Reliable DSS as proposed by DoD's AI guidelines (Board, 2019). The results also have implications for basic research in judgment and decision-making of how people's algorithm adoption and aversion behaviors manifest (Dietvorst & Bharti, 2020), particularly that people are able to adapt their decision strategies to the level of algorithmic misspecification. Also, we found support that people were using compensatory strategies instead of relying on heuristic strategies in our multi-objective tasks, consistent with previous findings within the Adaptive Decision Maker (Payne et al., 1993) as opposed to simply engaging in effort reduction via heuristic use (Shah & Oppenheimer, 2008).

5.2 Future Directions

As the next step, I plan to replicate this study at another university to test the difference in results between the two population samples. It will help better understand the differences in performance between conditions. The current experiments can also be

replicated in the future by using stronger misspecification and more attributes and alternatives so that corrective strategy and equal weights are more apart in their statistical estimation and more micro-level conclusions can be drawn. Future work can also extend the current experiments for multiple different conditions, including testing different mitigation techniques. It can evaluate the proposed mitigation technique in Experiment 2 by manipulating the quality of explanation, e.g., accurate complete explanation, accurate partial explanation, inaccurate complete explanation, inaccurate partial explanation, etc. The type of explanation can also be manipulated to see differences between people's understanding of model vs. outcome-based explanations. As discussed earlier, Páez (2019) and other XAI researchers usually recommend focusing more on model outcome-focused simplified explanations. However, the results of this study, and those from judgment and decision-making literature (Newell et al., 2009; Steinmann, 1976; Todd & Hammond, 1965), provide support for presenting more process-based feedback concerning the MDSS misspecification instead of outcome feedback to improve learning in a multiple-cue probabilistic environment. In this study, I presented participants with process feedback about how the model of MDSS is misspecified, and participants were found to adapt to that explanation. An alternative to this explanation could be outcome feedback that would provide feedback on the MDSS's performance or calibration level without explaining why the model is performing a certain way and why it might fail. This suggested explanation manipulation will help understand how the consequences of model blindness depend on the quality and type of mitigation and what can be the consequences of inaccurate mitigation. Future work can also test different types of model misspecification except for cue-weighting, e.g., missing vs. complete set of cues.

Future work should also focus on challenges posed by model blindness apart from performance degradation, such as system brittleness, trust issues affecting MDSS adoption behavior by users, and algorithm aversion (Dietvorst & Bharti, 2020). Auditing MDSS for its potential to cause model blindness is also a necessary next step. Through a series of multiple experiments, future work can quantify the framework and develop metrics of model blindness analogous to the one implemented in this paper (Figure 6) appropriate for various tasks that can generate signals of risk or performance degradation that an MDSS might impose on operators. Metrics like these can also assess the performance degradation that an MDSS might impose on operators using different decision strategies within the same context. Hence, providing a range of conditions under which an operator's performance under model blindness will be more or less likely to be model-limited.

APPENDIX A. INSTRUCTIONS PRESENTED TO PARTICIPANTS

Consent Form Page 1

Image of consent form

Click the "ACCEPT" button to proceed.

Instructions Page 2:

Welcome to the Decision Processes Lab!

Thank you for agreeing to participate in this study! To begin, you will first learn to play our Route Planning Game using an AI-based Route Recommender System.

You must read all of these instructions very carefully to develop a good understanding of what you will do during the game.

Please do not take any notes during the game.

Instructions Page 3 (Control Conditions):

Mission-Planning tasks for asset routing like the one you will do in today's experiment are prevalent in commercial and navy shipping maritime operations. There has been a lot of research on developing decision support tools to support these tasks.

Today, you have to take on the role of a remote route-planning operator. You will be responsible for planning a route for delivering critical care shipment for COVID-19

patients from one geographical location to another via a route that involves a majority fleet movement and some ground shipping. The game will start with a practice round where you learn how to perform the route-planning task.

To achieve this goal, you will be provided with a Route Recommender System, an artificially intelligent system that helps make decisions by presenting routes and relevant information associated with routes. The Route Recommender System presents a set of seven best routes (represented as solid yellow lines) that are identified as top routes by the system to meet task goals.

The goal is to choose the best route for every scenario out of the top seven routes presented by the Route Recommender System. All scenarios will have one best correct route out of the seven presented.

Instructions Page 3 (Experimental Recommender and Mitigation Conditions):

Mission-Planning tasks for asset routing like the one you will do in today's experiment are prevalent in commercial and navy shipping maritime operations. There has been a lot of research on developing decision support tools to support these tasks.

Today, you have to take on the role of a remote route-planning operator. You will be responsible for planning a route for delivering critical care shipment for COVID-19 patients from one geographical location to another via a route that involves a majority fleet movement and some ground shipping. The game will start with a practice round where you learn how to perform the route-planning task.

To achieve this goal, you will be provided with a Route Recommender System, an artificially intelligent system that helps make decisions by presenting routes and relevant information associated with routes. The Route Recommender System presents a set of three routes (represented as solid yellow lines) that are identified as top routes by the system to meet task goals. The system also provides a set of four additional routes (represented as dashed red lines) to choose from. You are free to pick any route you would like to.

The goal is to choose the best route for every scenario out of the top seven routes presented by the Route Recommender System. All scenarios will have one best correct route out of the seven presented.

Instructions Page 4

To help complete the route-selection task, you can review how each route scores on six attributes that are critical for decision-making in this task. The six attributes include time efficiency, fuel efficiency, obstacle avoidance (e.g., road closures, route deviations), weather hazard avoidance (flooded roads, hurricanes), en-route availability of extra supplies (vaccine shipments, medication shipments), and en-route ability to do humanitarian aid (deliver materials to people in need). The score on each attribute ranges from 0 to 100. A higher number means a particular route is better on a given attribute, e.g., a value of 90 on time efficiency attribute means the particular route is highly time-efficient, a value of 10 on weather avoidance attribute means the route is encountering a lot of hazards and not doing great in avoiding hazards.

You have to use all the presented attribute information. Each piece of information is equally important for an optimal decision. You have to optimize all attributes to meet multiple

objectives (e.g., maximize time efficiency while maximizing weather hazard avoidance) via your route selection.

You will also be provided ranked feedback about your chosen route after every choice.

Instructions Page 5

Try to do the best you can and utilize the feedback presented to you as much as you can.

You will not be given any information other than feedback to learn more about the relationship between route attributes and the best possible routes.

Instructions Page 6

You are now ready to play the practice round of the Route-Planning Game!

If you have any questions or are unsure of your task, ask the experimenter before beginning the game.

Instructions for the practice round (4 trials):

Click button (white squares) A-G (for routes A-G) to look at the attribute values for each route.

Make sure to select the route you believe is best, and press SUBMIT button to complete each scenario.

Instructions Page 7

Practice Round Ends here!

You are now ready to play the Route-Planning Game!

If you have any questions or are unsure of your task, ask the experimenter before beginning the game.

Instructions Page 8

Game Completed!

Good job picking the routes! You did great today!

Instructions Page 9: Debriefing 1

Thank you for your participation!

We are interested in investigating your decision-making when using the route information provided by a recommender system using an algorithm. Participants in different conditions of the experiment receive a recommender system with different levels of accuracy, affecting the quality of information they receive. We are trying to understand better how misspecifications and errors in algorithms can affect human decision-making.

The work has the potential to improve recommender systems used in high consequence decision-making tasks like medical diagnosis, navy ship routing, routing UAVs (Unmanned Aerial Vehicles), etc.

Instructions Page 10: Debriefing 2

Due to the sensitive nature of the research process, please do not discuss this procedure with other people that may participate in this experiment.

If you do not have any questions, you can gather your belongings and quietly leave the lab. If you have questions regarding the study, please talk to the experimenter available in the lab. You may also contact the experimenters via the email addresses listed on the SONA page for the experiment you used to sign up for the study.

If your participation in this study has upset you, please contact Georgia Tech's Office of Research Integrity Assurance.

THANK YOU AND GOODBYE!

APPENDIX B. PRE-STUDY DEMOGRAPHICS QUESTIONNAIRE

Question	Response format	Response Options
What is your age (in years)?	Free-text response	Textbox
What is your gender?	Forced-choice	Female, Male, Non-binary or Non-conforming, or Other, Prefer not to say
What is your major at Georgia Tech	Free-text response	Textbox
Do you have any experience playing computer-based video games?	Forced-choice	Yes, No
How often do you play computer-based video games?	Forced-choice	Very Often, Often, Occassionally, Rarely, Never

APPENDIX C. POST-STUDY FEEDBACK AND EXPERIENCE QUESTIONNAIRE

Question	Response format	Response Options
Please describe any limitations/issues you noticed with the AI-based route recommender system?	Free-text response	Textbox
How did the limitations you listed above (if any) affect your performance in selecting the best route in the experiment today?	Free-text response	Textbox
What additional information would you prefer to receive about/from the AI-based route recommender system?	Free-text response	Textbox
Please describe any limitations/issues you noticed with the "warning message" you received about the error in the AI-based route recommender system provided to you? *	Free-text response	Textbox
What additional information would you prefer to receive about the AI-based route recommender system via the "warning message"? *	Free-text response	Textbox
Do you have any experience using AI-based recommender systems (e.g., Netflix movie recommendations, Amazon product recommendations, Health and Fitness apps) in your day-to-day life?	Forced-choice	Yes, No
How often do you use recommendations provided by any AI-based recommender systems (e.g., Netflix movie recommendations, Amazon product recommendations, Health and Fitness apps) in your day-to-day life?	Forced-choice	Very Often, Often, Occassionally, Rarely, Never

Question	Response format	Response Options
Have you ever taken a course or worked with recommender systems?	Forced-choice	Yes, No

*Explanation conditions only

REFERENCES

Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675*.

Avvari, G. V., Sidoti, D., Zhang, L., Mishra, M., Pattipati, K., Sampson, C. R., & Hansen, J. (2018). Robust multi-objective asset routing in a dynamic and uncertain environment. 2018 IEEE Aerospace Conference,

Baeza-Yates, R. (2016). Data and algorithmic bias in the web. Proceedings of the 8th ACM Conference on Web Science,

Beemer, B. A., & Gregg, D. G. (2008). Advisory systems to support decision making. In *Handbook on Decision Support Systems 1* (pp. 511-527). Springer.

Ben-Haim, Y. (2006). *Info-gap decision theory: decisions under severe uncertainty*. Elsevier.

Board, D. I. (2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense: Supporting Document. *United States Department of Defense*.

Box, G. E., & Luceno, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment*. John Wiley & Sons, Inc.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3.

Brunswik, E. (1952). The conceptual framework of psychology. *Psychological bulletin*, *49*(6), 654-656.

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature human behaviour*, *5*(12), 1636-1642.

Cerutti, F., Kaplan, L. M., Kimmig, A., & Şensoy, M. (2022). Handling epistemic and aleatory uncertainties in probabilistic circuits. *Machine Learning*, 1-43.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (Army Research Lab Aberdeen, Issue.

Dhaliwal, J. S., & Benbasat, I. (1996). The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation. *Information systems research*, *7*(3), 342-362.

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment*, *19*(2), 209-216.

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.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269-271.

Dodson, T., Mattei, N., Guerin, J. T., & Goldsmith, J. (2013). An English-language argumentation interface for explanation generation with Markov decision processes in the domain of academic advising. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *3*(3), 1-30.

Doyen, S., & Dadario, N. B. (2022). 12 Plagues of AI in Healthcare: A Practical Guide to Current Issues With Using Machine Learning in a Medical Context. *Frontiers in Digital Health*, *4*.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), 697-718.

Dzindolet, M. T., Pierce, L., Pomranky, R., Peterson, S., & Beck, H. (2001). Automation reliance on a combat identification system. Proceedings of the Human Factors and Ergonomics Society Annual Meeting,

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open*, *9*(1), 2158244019832705.

Ehsan, U., & Riedl, M. (2019). On design and evaluation of human-centered explainable AI systems. *Glasgow'19*.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human factors*, *37*(1), 32-64.

Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. *Fox, Craig R. and Gülden Ülkümen (2011),"Distinguishing Two Dimensions of Uncertainty," in Essays in Judgment and Decision Making, Brun, W., Kirkebøen, G. and Montgomery, H., eds. Oslo: Universitetsforlaget*.

Gittins, R. (2012). Using psychology to improve economic policy. *Australian Economic Review*, *45*(3), 379-385.

Heuer Jr, R. J., & Pherson, R. H. (2010). *Structured Analytic Techniques for Intelligence Analysis*. CQ Press.

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, *54*(2-3), 217-223.

Illingworth, D. A., & Feigh, K. M. (2021). Impact Mapping for Geospatial Reasoning and Decision Making. *Human factors*, 0018720821999021.

Jameson, A., Willemsen, M. C., Felfernig, A., de Gemmis, M., Lops, P., Semeraro, G., & Chen, L. (2015). Human decision making and recommender systems. In *Recommender systems handbook* (pp. 611-648). Springer.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71.

Johnson, J. D., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Type of automation failure: The effects on trust and reliance in automation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting,

Kizzier-Carnahan, V., Artis, K. A., Mohan, V., & Gold, J. A. (2019). Frequency of passive EHR Alerts in the ICU: another form of alert fatigue? *Journal of patient safety*, *15*(3), 246.

Kurz-Milcke, E., & Gigerenzer, G. (2007). Heuristic decision making. *Marketing: Journal of Research and Management*, *3*(1), 48-56.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lawrence, A., Thomas, R. P., & Dougherty, M. R. (2018). Integrating Fast and Frugal Heuristics with a Model of Memory-based Cue Generation. *Journal of Behavioral Decision Making*, *31*(4), 487-507.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Liao, T. F., & Fasang, A. E. (2021). Comparing Groups of Life-Course Sequences Using the Bayesian Information Criterion and the Likelihood-Ratio Test. *Sociological Methodology*, *51*(1), 44-85. https://doi.org/10.1177/0081175020959401

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of consumer research*, *46*(4), 629-650.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, *66*(2), 81.

Mangiameli, P., West, D., & Rampal, R. (2004). Model selection for medical diagnosis decision support systems. *Decision support systems*, *36*(3), 247-259.

Marakas, G. M. (2003). *Decision support systems in the 21st century* (Vol. 134). Prentice Hall Upper Saddle River, NJ.

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors*, *58*(3), 401-415.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

Mosier, K. L., & Fischer, U. M. (2010). Judgment and decision making by individuals and teams: issues, models, and applications. *Reviews of human factors and ergonomics*, *6*(1), 198-256.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905-1922.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429-460.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *26*(1), 41-47.

Newell, B. R., Weston, N. J., Tunney, R. J., & Shanks, D. R. (2009). The effectiveness of feedback in multiple-cue probability learning. *Quarterly Journal of Experimental Psychology*, *62*(5), 890-908.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive psychology*, *7*(1), 44-64.

O'Neil, C. (2016). *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy* (Vol. First edition) [Book]. Broadway Books.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441-459.

Papamichail, K. N., & French, S. (2005). Design and evaluation of an intelligent decision support system for nuclear emergencies. *Decision support systems*, *41*(1), 84-111.

Parmar, S., Illingworth, D. A., & Thomas, R. P. (2021). Model blindness: A Framework for Understanding how Model-Based Decision Support Systems can Lead to

Performance Degradation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting,

Parmar, S., & Thomas, R. P. (2020). Effects of Probabilistic Risk Situation Awareness Tool (RSAT) on Aeronautical Weather-Hazard Decision Making. *Frontiers in psychology*, *11*.

Payne, J. W. (1980). Information processing theory: Some concepts and methods applied to decision research. *Cognitive processes in choice and decision behavior*, *95*, 115.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge university press.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), 195-203.

Peirce, J., Hirst, R., & MacAskill, M. (2022). *Building experiments in PsychoPy*. Sage.

Power, D. J., & Sharda, R. (2007). Model-driven decision support systems: Concepts and research directions. *Decision support systems*, *43*(3), 1044-1061.

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455-468.

Rebitschek, F. G., Gigerenzer, G., & Wagner, G. G. (2021). People underestimate the errors by algorithms for credit scoring and recidivism but tolerate even fewer errors.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychological bulletin*, *134*(2), 207.

Shanteau, J., & Thomas, R. P. (2000). Fast and frugal heuristics: What about unfriendly environments? *Behavioral and Brain Sciences*, *23*(5), 762-763.

Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H., & Tang, P. C. (2001). Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association*, *8*(6), 527-534.

Steinmann, D. O. (1976). The effects of cognitive feedback and task complexity in multiple-cue probability learning. *Organizational Behavior and Human Performance*, *15*(2), 168-179.

Stuck, R. E., Tomlinson, B. J., & Walker, B. N. (2022). The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science*, *23*(4), 500-516.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, *3*(1), 1-10.

Todd, F. J., & Hammond, K. R. (1965). Differential feedback in two multiple-cue probability learning tasks. *Behavioral science*, *10*(4), 429-435.

Turban, E., Aronson, J. E., & Ting-Peng, L. (2001). Decision support systems and intelligent systems, 6th Prentice-Hall International. *Upper Saddle River (NJ, USA)*.

van Leeuwen, C., Smets, A., Jacobs, A., & Ballon, P. (2021). Blind spots in AI: the role of serendipity and equity in algorithm-based decision-making. *ACM SIGKDD Explorations Newsletter*, *23*(1), 42-49.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology and human performance*. Psychology Press.

Woods, D. D. (1993). Price of flexibility in intelligent interfaces. *Knowledge-Based Systems*, *6*(4), 189-195. https://doi.org/https://doi.org/10.1016/0950-7051(93)90011-H

Woods, D. D., & Cook, R. I. (2006). Incidents–markers of resilience or brittleness. *Resilience engineering: Concepts and precepts*, 69-76.

Woods, D. D., Patterson, E. S., & Roth, E. M. (2002). Can we ever escape from data overload? A cognitive systems diagnosis. *Cognition, Technology & Work*, *4*(1), 22-36.

Woods, D. D., & Sarter, N. B. (1998). Learning from automation surprises and" going sour" accidents: Progress on human-centered automation. *NASA*(19980016965).

Yates, J. F. (1990). *Judgment and decision making*. Prentice-Hall, Inc.