ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240571

Towards the New XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence

Thao Le[a,*], Tim Miller[b], Liz Sonenberg[a] and Ronal Singh[c]

[a]School of Computing and Information Systems, The University of Melbourne
[b]School of Electrical Engineering and Computer Science, The University of Queensland
[c]CSIRO's Data61

Abstract. Prior research on AI-assisted human decision-making has explored several different explainable AI (XAI) approaches. A recent paper has proposed a paradigm shift calling for *hypothesis-driven XAI* through a conceptual framework called *evaluative AI* that gives people evidence that supports or refutes hypotheses without necessarily giving a decision-aid recommendation. In this paper, we describe and evaluate an approach for hypothesis-driven XAI based on the Weight of Evidence (WoE) framework, which generates both positive and negative evidence for a given hypothesis. Through human behavioural experiments, we show that our hypothesis-driven approach increases decision accuracy and reduces reliance compared to a recommendation-driven approach and an AI-explanation-only baseline, but with a small increase in under-reliance compared to the recommendation-driven approach. Further, we show that participants used our hypothesis-driven approach in a materially different way to the two baselines.

1 Introduction

Research has shown that AI recommendations, even when accompanied with explanations, are not always helpful in supporting decision-making [4, 38, 6, 15]. The direct causes of this are *under-reliance* and *over-reliance* [37]. With under-reliance, decision-makers reject AI recommendations, even when they may be correct. Alternatively, decision-makers may overly rely on AI recommendations, hence be led to errors when the AI is incorrect. In either case, they tend to fixate on a particular hypothesis without sufficiently considering others [27]. Approaches such as *cognitive forcing*, based on ideas from human psychology, have been proposed to address limitations of the AI recommendation approach [6], with recent work indicting that withholding AI model the recommendations, at least for a short time, while still providing the user with an explanation of that recommendation, can be helpful [15]. Recently, Miller [27] proposed a so-called **hypothesis-driven** decision-making paradigm called **evaluative AI**. The main aim of this paradigm is to focus the decision loop on the human decision maker, providing them with the right evidence to support their own intuitions, rather than focusing the decision loop on machine recommendations. This paradigm offers a promising direction in building better decision support in explainable AI (XAI) research by focusing on human decision makers considering multiple possible hypotheses.

* Corresponding Author. Email: thao.le@unimelb.edu.au

In this paper, we describe and evaluate an approach for building a hypothesis-driven decision-making model that uses the *Weight of Evidence (WoE)* framework [2]. To the best of our knowledge, this is the first work that compares empirically and in a controlled manner the hypothesis-driven approach [27] with two other popular decision-making approaches (recommendation-driven and AI-explanation-only [15]). Our contributions are:

- The *Evidence-Informed Hypothesis-Driven Decision-Making* model, building on the *Weight of Evidence (WoE)* framework to the hypothesis-driven approach;
- Two human behavioural experiments comparing our *hypothesis-driven* approach with two common decision-aid approaches: (1) the standard model recommendation with explanation; and (2) a form of cognitive forcing that provides only AI explanations [15]. The results show that hypothesis-driven approach improved decision accuracy and reduces over-reliance compared to standard recommendation-driven approaches, at the cost of a slight increase in under-reliance. Furthermore, the hypothesis-driven approach reduces under-reliance significantly compared to the AI-explanation-only approach. Our qualitative analysis identifies some limitations and challenges in the three approaches, and shows that participants used the hypothesis-driven approach in a materially different way than the recommendation-driven or AI-explanation-only conditions, with participants focusing more on the evidence than on their own background knowledge.

2 Background and Related Work

In this section, we give a brief overview of related work on common AI-assisted decision-making paradigms. We then highlight key literature on applying explainable AI in supporting decision-making.

2.1 AI-Assisted Decision-Making Paradigms

In the literature, there are two workflows that are often used in AI-assisted decision-making: (1) AI-first decision-making; and (2) human-first decision-making. AI-first workflow provides the AI recommendation first and then humans decide if they want to accept or not the recommendation, whereas human-first decision-making requires humans to make a provisional decision before they are provided with any AI recommendations.

AI-first Workflow In the AI-first decision-making workflow, it has been demonstrated that participants feel more confident and are also faster in decision-making [14]. These participants also rated AI as more practical. In terms of limitations, the anchoring effect is reported to occur more often in the AI-first workflow [14, 29, 6] in which people overly rely on the AI recommendation (also called *over-reliance*). The anchoring effect [34] refers to giving stronger preference to the earlier knowledge rather than doing a full revision and considering the latest evidence. By contrast, Fogliato et al. [13] did not find any significant difference in the participants' performance, which is measured by accuracy between the two workflows (AI-first and human-first). However, they also found that participants are 65% more likely to revise their answers in the human-first setting than those in the AI-first setting.

Human-first Workflow Human-first decision workflow has been shown to help reduce the over-reliance on erroneous AI recommendations [6, 14]. However, experts may interact with decision-making systems differently from laypeople (crowdworkers). For example, Fogliato et al. [14], Gaube et al. [16] run studies with radiologists who were the experts. Their task is to review patients' X-ray images. The studies conclude that in human-first workflow with expert participants, they are less likely to leverage AI advice even though the AI is more accurate. In fact, this is referred to *algorithm aversion* [12] or *under-reliance*.

A human-first paradigm called *cognitive forcing*, based on earlier ideas in psychology for interventions that elicit human thinking at decision-making time [22], has been proposed as a way to improve users' engagement and also increase their learning when interacting with the AI [6, 15]. Four cognitive forcing designs have been introduced: (1) *On demand*: Participants can only see the AI recommendation when they request it; (2) *Update*: Participants first made a decision without seeing the AI recommendation. Then, they were shown the AI prediction and could update their decision later; (3) *Wait*: Participants had to wait for 30 seconds before the AI decision was shown; and (4) *Only AI explanation*: Providing just the AI explanation and *no AI recommendation*, on the basis that this may help people process the AI explanation more carefully and therefore, improve their knowledge and make better decisions [15]. Importantly, *cognitive forcing* has been shown to reduce *over-reliance* compared to the standard AI suggestion approach, although that study had a limitation in that the AI prediction was always correct. There are also some recognised trade-offs of cognitive forcing designs: more time-consuming [15], and less trust [6].

Evaluative AI (Hypothesis-driven) Paradigm Miller [27] argues that AI-assisted decision support is on the cusp of a paradigm shift. This shift is away from the idea of human-first or AI-first, and into a framework he calls **evaluative AI**. The key insight of evaluative AI is to not necessarily provide a recommendation, and instead to support the human cognitive decision-making process by providing evidence for or against the particular hypothesis that a human decision maker is considering. Miller argues that this would help to prevent over- and under-reliance, and would help the decision maker to retain their *internal locus on control* [31].

2.2 *Explainable AI (XAI) in Decision Support*

In this section, we review some AI-first explainable AI approaches that have been used to provide explanations for the AI recommendation based on feature analysis. We will also discuss some popular evidence-based explanations that have been used in supporting decision-making.

Human decision-making reliance on AI support There is no straightforward position regarding when humans are well calibrated to accept AI-generated advice [37]. Overall, study participants appear more likely to accept an AI's recommendation when provided with explanations, regardless of the model's correctness [4, 18, 35]. Explanations can increase the accuracy of the human-AI team when the AI is correct, but *decrease* it when it is wrong, resulting in over-reliance on the AI's recommendations. Arguably, this is because the current explanation styles do not provide details of the underlying rationale of the AI model behaviour [35]. We therefore should be careful when selecting the explanation type as it can have a significant effect on whether users decide to rely on them [7, 25].

Evidence-based Explanations In this paper, we will generate *evidence-based explanations*, which are similar to feature importance explanations. The main difference is that *Weight of Evidence* uses log likelihoods and log odds ratios to generate explanations, whereas LIME [30] and SHAP [26] find feature importance by modifying the predictive posterior probability in various ways.

Evidence-based explanations have been applied to support decision-making and debug models in several prior research [19, 20, 28] other than from Alvarez Melis et al. [2]. A closely related work to Alvarez Melis et al. [2] is from Poulin et al. [28]. Poulin et al. [28] propose a framework called *ExplainD* that uses *additive evidence*. The framework also measures the weight of evidence using a Naive Bayes classifier along with highlighting the negative and positive evidence for a decision. However, the problem being considered is a binary classification. Furthermore, there is still room for improvement by conducting experiments to evaluate the framework. Kulesza et al. [19, 20] introduce *EluciDebug* in email classification using Multinomial Naive Bayes classifier (MNB). The EluciDebug prototype provides an interface that includes important words and the folder size that both contribute to the email classification. However, the prototype did not specifically give positive and negative evidence in decision-making situations.

3 Evidence-Informed Hypothesis-Driven Decision-Making Model

We define the *evidence-informed hypothesis-driven decision-making* model by implementing the *evaluative AI* (hypothesis-driven) paradigm [27] using the WoE model. Specifically, given a classification problem, decision makers explore evidence for and against each hypothesis (i.e. an output class). We allow decision-makers to interact with the model by repeatedly selecting a hypothesis for which they can then see the positive (or negative) evidence.

Evidence Generation In a classification problem, a hypothesis $h \in Y$, where $Y = \{h, h_1, h_2, ..., h_n\}$ includes all possible hypotheses, as an output class. Then, $\bar{h} = Y \setminus \{h\}$ refers to all hypotheses other than h. For example, if a doctor asserts a set of hypotheses $Y = \{h_1, h_2, h_3\}$ where h_1 = *the patient has Covid*, h_2 = *the patient has Influenza* and h_3 = *the patient has pneumonia*, then $\bar{h_1}$ = *the patient does not have Covid* which includes all possible hypotheses except having Covid, that is $\bar{h_1} = \{h_2, h_3\}$.

We generate the *weight of evidence* for possible hypotheses using Weight of evidence (WoE), which is a probabilistic approach for analysing variable importance, introduced in the context of explainability by Alvarez Melis et al. [2] building on the approach of Good [17]. It provides a quantitative response to the question of why a model predicted output h for a particular input x in terms of how much each input feature x_i provides in favour of, or against, h, relative to alternatives. Through Bayes rule, WoE can be understood as an adjustment to the prior log odds caused by observing the evidence.

For hypothesis h and input feature x_i, weight of evidence, woe, is defined as follows:

$$\mathrm{woe}(h \mid x_i) = \log \frac{P(x_i \mid h)}{P(x_i \mid \bar{h})} = \log \frac{P(h \mid x_i)}{P(\bar{h} \mid x_i)} - \log \frac{P(h)}{P(\bar{h})} \quad (1)$$

Based on the weight of evidence, we say the evidence supports or refutes a hypothesis:

- If $\mathrm{woe}(h \mid x_i) > 0$, evidence x_i supports hypothesis h
- If $\mathrm{woe}(h \mid x_i) < 0$, evidence x_i refutes hypothesis h
- If $\mathrm{woe}(h \mid x_i) = 0$, evidence x_i neither supports or refutes hypothesis h

How decision-aid models can use WoE to make a decision Using the weight of evidence for each feature x_i as in Equation 1, a decision-aid model can make a prediction based on the total weight of evidence of a hypothesis h by summing up the weight of evidence of this hypothesis based on each feature x_i. The total weight of evidence is defined as follows.

$$\mathrm{woe}(h) = \sum_{i=1}^{n} \mathrm{woe}(h \mid x_i) \quad (2)$$

where n is the number of features.

The decision-aid model will select the best hypothesis based on the maximum posterior, that is, $y = \arg\max_{h \in Y} P(h \mid X)$. If we have the same prior for all hypotheses, we can also use the total weight of evidence as another way to find the best hypothesis using Equation 1. Therefore, a decision-aid model can select the hypothesis with the maximum total weight of evidence as its prediction as follows (only apply to uniform priors).

$$y = \arg\max_{h \in Y} \mathrm{woe}(h) \quad (3)$$

How WoE can incorporate a human approach to making a decision To assist users with interpretability, Alvarez Melis et al. [2] complement the display of the magnitude of the weight of the evidence with a notion of significance level of the evidence, using a scale of seven categories: *decisive-against (---), strong-against (--), substantial-against (-), not-significant (N), substantial-in-favor (+), strong-in-favor (++), decisive-in-favor (+++)*. The details can be found in the rule-of-thumb guidelines here [3].

In addition to the weight of evidence of a feature, we suggest it is useful to distinguish the *importance* of a feature – with importance being domain specific and determined by the domain expert using the model. Specifically, if a feature has significant weight of evidence according to the WoE model, but that feature is not seen as important by the human decision maker, then it is reasonable to anticipate the impact of that evidence on the decision would be reduced by the decision maker. For example, if a clinician looks at a dermatoscopic image and also is aware of some irrelevant but high weight of evidence feature such as dense hair, they should ignore that evidence in making a prediction.

Formally, by considering the importance of the evidence, we redefine the total weight of evidence from a human decision-making perspective as follows:

$$\mathrm{woe}(h) = \sum_{i=1}^{n} \gamma_i \times \mathrm{woe}(h \mid x_i) \quad (4)$$

where γ_i is a parameter of feature x_i that adjusts the weight of evidence based on importance, i.e., $\gamma_i > \gamma_j$ represents that feature x_i is more important than feature x_j.

Then, for the skin cancer example just mentioned, in effect in Equation 4, the clinician has set $\gamma = 0$ for that feature.

4 Experiment Design

In this section, we describe the task implemented in the human behaviour experiment and the experiment design. In selecting a decision-making task, we identified requirements similar to those used in other studies of how explanations can assist human decision makers interacting with AI decision support, e.g. Vasconcelos et al. [36]: the task should not be too easy for humans to complete without a decision aid, but also, as we were using lay subjects from Prolific for this particular study, the task cannot require specialist knowledge.

We chose a version of the *housing price prediction* task studied previously in an XAI context [1, 10]. In this task, participants are provided with information about house features and with other information which varies by experimental condition, and are asked to choose whether the given house would have a sale price of *low*, *medium* or *high*. As noted by others [10], real estate valuation is a domain where ML models have been developed to help people make better decisions, predicting house prices is a task that lay people may need to do in real life, so it is not unrealistic to expect they have sufficient day-to-day knowledge to make predictions and decide whether or not to rely on an AI model.

Experimenting with this task, we compared the *hypothesis-driven* approach with two state-of-the-art decision-making approaches using quantitative measures for *efficiency*, *performance* and *reliance* and a qualitative analysis of *information use*. In the terminology of a recent review of XAI evaluation [21], the first two points of comparison are a form of evaluation with respect to the decision task, and the latter two focus on users' perception and use of the AI system itself.

4.1 Dataset and Model Implementation

To build our model, we used the Ames Housing Dataset [11] and the open source code on GitHub [33] for data pre-processing. The data after pre-processing has a total of 2616 instances and 28 features. We processed the dataset further by converting the house price into three output classes (*low price*, *medium price* and *high price*). We also balanced the dataset to ensure that the three classes had the same number of instances by using Near-Miss Undersampling. Finally, we had a total of 1920 instances with 640 instances for each class.

We selected six features for the human experiment in the house-price decision-making task by applying Gradient Boosting Classification model over the data. Considering domain specific decision-making about house prices, we propose there to be three important features (*quality of construction*, *house age* and *location*) and three unimportant features (*fireplaces*, *kitchen quality* and *central air conditioning*). We divided the dataset into 80% for the training set and 20% for the test set. Following Alvarez Melis et al. [2], we use a Gaussian Naïve Bayes (GNB) classifier to obtain $P(x_i \mid h)$. This assumes that features are independent, but the model and implementation work for any probabilistic classifier. We chose this model because it is a simple discriminative classifier that aligns with previous work on evidence-based explanations [28, 20].

4.2 Experimental Conditions

All participants[1] were given the six house feature values plus other information, which varied by condition as set out below. Participants then chose whether the given house would have a price of *low*, *medium* or *high*. Using a *between-subject design*, participants were randomly assigned to one of three conditions:

- (C1) *Recommendation-driven*: Participants see the AI prediction (i.e., either *low* or *medium* or *high*) *and* also the weight of evidence for that prediction;
- (C2) *AI-explanation-only*: Participants see the weight of evidence associated with the AI prediction, but the AI prediction itself is hidden;
- (C3) *Hypothesis-driven*: Participants see the weight of evidence for *all* hypotheses (*low*, *medium* and *high*), but the AI prediction itself is hidden.

Although participants in the *AI-explanation-only* and *hypothesis-driven* conditions did not see a recommendation, it was expected that the displayed information from the WoE framework would provide insight that participants could use to support their decision-making. We note a similar AI-explanation-only approach has been explored previously [15].

4.3 Research Questions and Hypotheses

Our overarching research questions were as follows:

- **RQ1**: *(Efficiency)* What form of AI assistance helps participants make faster decisions?
- **RQ2**: *(Performance)* What form of AI assistance helps participants make better decisions?
- **RQ3**: *(Reliance)* What form of AI assistance helps reduce over-reliance and under-reliance?
- **RQ4**: *(Information use)* How do people make decisions differently in *recommendation-driven*, *AI-explanation-only* and *hypothesis-driven* paradigm?

For **RQ1**, we evaluated the participants' speed in making a decision. We use the most common metric - *completion time* to measure the time taken on the task. The corresponding hypotheses for this question are:

- **H1a/b**: *(C3) Hypothesis-driven* paradigm will cost less time to finish the task than *(C1) Recommendation-driven* and *(C2) AI-explanation-only*.

For **RQ2**, we evaluated the quality of the decision. In the task, we asked the participants to assign the likelihood for each price range (low/medium/high) where 100 is the most likely and 0 is the least likely. The sum of three likelihoods must be equal to 100. We expect the participants to be confident when they make a correct prediction, and *not be confident when they make a wrong decision*. We apply *Brier score* as explained below to measure the task performance. The hypotheses for this question are:

- **H2a/b**: *(C3) Hypothesis-driven* paradigm will help participants make better decisions than *(C1) Recommendation-driven* and *(C2) AI-explanation-only*.

[1] We received ethics approval from our institution before conducting the human experiment.

For **RQ3**, we investigated the participants' capability of appropriately calibrating their decision. Participants should follow the model's prediction when it is correct and should not use the model's prediction when it is wrong. We applied two measures *over-reliance* and *under-reliance* as shown below with the following hypotheses:

- **H3a**: *(C3) Hypothesis-driven* can reduce over-reliance compared to *(C1) Recommendation-driven*.
- **H3b**: *(C3) Hypothesis-driven* can reduce under-reliance compared to *(C2) AI-explanation-only*.

For **RQ4**, we looked into the text written by participants when they explained why they selected an option after each question to know how they used the provided information in each decision-making paradigm to make their decisions. Therefore, we can identify the limitations of each paradigm and the generated evidence that lead the participants to make a wrong decision.

4.4 Measures

We took the following measures:

1. *Task Efficiency* (Completion time): The time participants take to complete each task.
2. *Task Performance* (Brier score): This metric quantifies the effectiveness of task performance in terms of accurate decision outcomes. The formula is:

$$\mathrm{BS}_p = \frac{1}{N}\sum_{i=1}^{N}(C_{p,i} - A_{p,i})^2 \qquad (5)$$

where: $C_{p,i}$ is the likelihood level of participant p in question i, ranging from 0 to 1; $A_{p,i}$ is the answer score of participant p in question i, either 0 (wrong answer) or 1 (right answer); N is the number of questions for each participant. The best Brier score (i.e. equal to 0) for an individual task is when a participant answers the task correctly and gives it a 100% likelihood (or alternatively, a wrong answer but with 0% likelihoood). Therefore, a participant has better task performance when they have a lower Brier score. The Brier score measures decision accuracy, but awards a higher score for a correct answer when a participant is confident, and lower penalty for an incorrect answer when not confident. This mitigates problems where participants guess answers (i.e. have low confidence in their answers).

We then measure the *over-reliance* and *under-reliance* [38]. Since study participants can only see the AI recommendation in C1 (*Recommendation-driven*), we measure *agreement*: whether participants have the same prediction or differ from the model's prediction in the other two conditions.

3. *Over-reliance*: the fraction of tasks where participants have the same decision as a model's prediction when it was wrong: $\Sigma_i(A_{p,i} = M_i = 0)/\Sigma_i(1 - M_i)$, where $A_{p,i}$ is as above and $M_i = 1$ if the model is correct and 0 otherwise.
4. *Under-reliance*: the fraction of tasks where participants have a different decision from a model's prediction when it was correct: $\Sigma_i(A_{p,i} \neq M_i = 1)/\Sigma_i M_i$.

4.5 Conduct

We conducted ***two*** separate human experiments in which participants were given the same task in the form of a question set, with the only difference being the way they answered the question.

In experiment 1, participants were asked to make a decision about the relative likelihood for each price range (*low*/*medium*/*high*) of given house instances. We answer **RQ1**, **RQ2** and **RQ3** by analysing the results of four measures mentioned above (completion time, Brier score, over-reliance and under-reliance).

In experiment 2, we recruited a new and smaller cohort and asked them to do the same tasks as in experiment 1, but in addition, we asked participants to explain their decisions using free text. We conduct this experiment separately from the quantitative data in experiment 1 because asking participants to explain their reasoning cognitively forces them to engage with the instance, interfering with their natural decision-making process, and therefore potentially affecting the quantitative results. We then performed a deductive analysis of their explanations to answer **RQ4**.

Each experiment was designed as a Qualtrics[2] survey and participants accessed the survey through Prolific[3]. The experiment required a maximum of 25 minutes to finish. There were 12 house instances given, equivalent to 12 questions. These 12 questions were evenly distributed into four question categories: (1) where the model gives *correct* predictions with *high uncertainty*, (2) where the model gives *correct* predictions with *low uncertainty*, (3) where the model gives *wrong* predictions with *high uncertainty* and (4) where the model gives *wrong* predictions with *low uncertainty*. Thus, there are three questions in each category. Study participants were *not* informed about in which category the question belongs. The questions are ordered randomly in the experiments.

The uncertainty is measured by the cross entropy as follows.

$$u(h) = -\sum_{h \in H} p(h) \log p(h) \tag{6}$$

where $u(h)$ is the uncertainty level of hypothesis h given the probabilistic output is $p(h)$. To ensure there was a clear difference between high and low uncertainty, we select instances with *low uncertainty* by choosing instances entropy less than 0.3, and *high uncertainty* by choosing instances with entropy greater than 0.7. Participants did not know the certainty nor how many test instances were correct/incorrect. Each participant was paid a minimum of £4 for their time, plus a bonus of £2 if they could answer at least 9 out of 12 questions correctly. Participants were also given a plain language statement, and consent form and did a training phase with 3 example questions before answering the 12 test questions.

For **RQ4** the text is a response to "Can you please explain why you selected this option?". We analysed a total of 12 (questions) × 95 (participants) = 1140 responses. The final analysis includes 1,031 responses after removing 109 responses due to poor quality. Each response is assigned to at least one category (or code): *Using feature values* or *Using evidence*. We then perform a simple deductive analysis by reading each response and assign the relevant codes. We explain each code as follows: (1) *Using Feature Values*: participants rely on the feature values and their background knowledge to make the final decision without using the model evidence; and (2) *Using Evidence*: Participants rely on the evidence provided by the model and possibly their background knowledge to make the final decision.

We chose these two codes based on the idea of *machine explanation* and *human intuition* [8, 9]. Specifically, *using evidence* refers to using the machine explanation and therefore, making use of the model evidence to support decision-making. On the other hand, *using feature values* is relevant to using people's intuitions of the task based on the input feature values. Therefore, using the qualitative analysis, we explore how people use the model evidence and their intuitions in the three decision-making paradigms.

4.6 Participants

Experiment 1 Using the power analysis for F-test for one factor ANOVA and assuming the power of 0.8 and significant alpha of 0.05, we found that a sample size of 300 participants in three groups guarantees a small effect size of 0.2. In total, we recruited $N = 302$ participants on Prolific, distributed into three conditions: 102 participants in C1, 99 participants in C2 and 101 participants in C3. Participants are selected from the United States, United Kingdom, and Australia and must be fluent in English. Gender-wise, 192 were women, 103 were men, 4 self-specified their gender and 3 declined to state their gender. Age-wise, 94 participants were between Ages 18 and 29, 91 were between Ages 30 and 39, 44 were between Ages 40 and 49, and 73 were over Age 50.

Experiment 2 We recruited $N = 95$ participants on Prolific, distributed into three conditions: 30 participants in C1, 34 participants in C2 and 31 participants in C3. Participants are selected from the United States, United Kingdom, and Australia and must be fluent in English. Gender-wise, 52 were women, 41 were men, and 2 declined to state the gender. Age-wise, 38 participants were between Ages 18 and 29, 37 were between Ages 30 and 39, 10 were between Ages 40 and 49, and 10 were over Age 50. Participants in the first study were *not* allowed to participate in the second.

5 Experiment Results

In this section, we show the results of two experiments. In the first experiment, we explore whether *hypothesis-driven* can improve task efficiency, task performance and reduce reliance compared to *recommendation-driven* and *AI explanation only*. In the second experiment, we understand how participants used our hypothesis-driven approach differently compared to the other two baselines. The full statistics and screenshots from our user study can be found in the supplementary material [24].

5.1 Experiment 1: Quantitative Results

We performed a Shapiro-Wilks test to check the data normality and we found that our data was not normally distributed ($p < 0.05$). Therefore, we apply non-parametric Kruskal-Wallis test. We then perform post-hoc Mann-Witney U test to do pairwise comparisons. The results are visualised in Table 1 and Table 1. The significant differences between two conditions are highlighted in italic red in the figures where $p < 0.05$.

Task efficiency Table 1 shows the completion time in three conditions. There is no statistically significant difference among these three conditions ($p \approx 0.9$). We reject **H1a/b**. ***This shows that hypothesis-driven does not take more time to complete the task than recommendation-driven and AI-explanation-only***.

Task performance We evaluate participants' decision-making performance by using the Brier score. A lower Brier score indicates better decision accuracy. As seen in Table 1, ***participants in the hypothesis-driven condition*** ($M = 0.267$, $SD = 0.063$) ***have a lower Brier score than the other two approaches*** (C1: ($M = 0.290$, $SD = 0.071$), C2: ($M = 0.295$, $SD = 0.073$)). We **accept H2a/b**. Therefore, hypothesis-driven helps participants be confident when

[2] https://www.qualtrics.com
[3] https://www.prolific.com

	R	O	H	R vs. O	R vs. H	O vs. H
Time (minutes)	$\mathbf{17.92 \pm 9.34}$	18.62 ± 11.26	18.09 ± 9.26	$p = 0.808, r = 0.020$	$p = 0.993, r = 0.001$	$p = 0.892, r = 0.011$
Brier score	0.29 ± 0.07	0.30 ± 0.07	$\mathbf{0.27 \pm 0.06}$	$p = 0.566, r = 0.047$	$\mathbf{p = 0.016, r = 0.196}$	$\mathbf{p = 0.003, r = 0.245}$
Over-reliance (%)	73.86 ± 20.91	54.21 ± 22.51	$\mathbf{53.30 \pm 22.73}$	$\mathbf{p < 0.001, r = 0.450}$	$\mathbf{p < 0.001, r = 0.449}$	$p = 0.946, r = 0.006$
Under-reliance (%)	$\mathbf{17.81 \pm 20.35}$	41.25 ± 27.18	24.42 ± 18.19	$\mathbf{p < 0.001, r = 0.523}$	$\mathbf{p = 0.001, r = 0.246}$	$\mathbf{p < 0.001, r = 0.387}$

Table 1. Results of the human experiment. R: Recommendation-driven, O: AI-explanation-only, H: Hypothesis-driven. Winners/significances are highlighted.

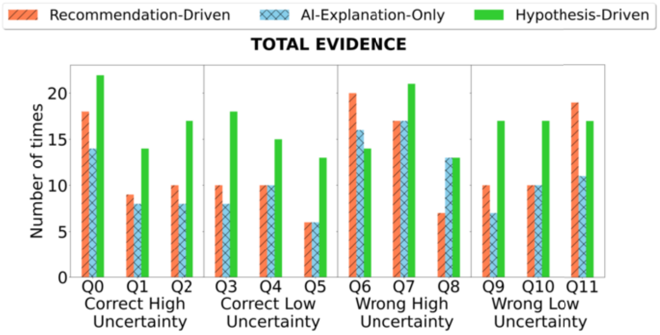

Figure 1. Frequency of using evidence to make a decision.

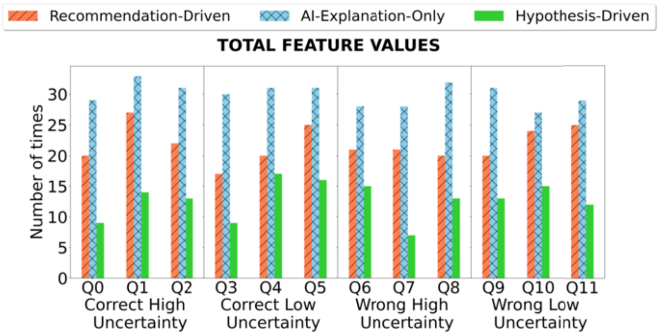

Figure 2. Frequency of using feature values to make a decision.

they make a correct decision, and be less confident when they make a wrong decision.

Over-reliance In Table 1, ***hypothesis-driven** (*$M = 53.30$*,* $SD = 22.73$*)* ***reduced over-reliance significantly compared to recommendation-driven*** *(*$M = 73.86$*,* $SD = 20.91$*)* ($p = 1.5 \times 10^{-8}$, $r = 0.449$). We **accept H3a**. Moreover, *AI-explanation-only* ($M = 54.21$, $SD = 22.51$) also reduces over-reliance compared to the recommendation-driven approach ($p = 1.6 \times 10^{-8}$, $r = 0.450$).

Under-reliance In Table 1, ***hypothesis-driven** (*$M = 24.42$*,* $SD = 18.19$*)* ***significantly reduced under-reliance compared to AI-explanation-only*** *(*$M = 41.25$*,* $SD = 27.18$*)* ($p = 1.09 \times 10^{-6}$, $r = 0.387$). We **accept H3b**. This is not surprising because we expect that participants in the *AI-explanation-only* condition are the most likely to 'under-rely' due to being unable to compare evidence across hypotheses nor see a recommendation. Recommendation-driven ($M = 17.81$, $SD = 20.35$) has the least under-reliance value because participants were given recommendations.

5.2 *Experiment 2: Qualitative Results*

In Figure 1 and 2, we illustrate the number of times that participants used feature values and evidence to make their decisions based on the text analysis.

For the recommendation-driven paradigm, participants use the feature values to confirm whether the decision aid's prediction and explanation are reliable or not. If participants think the feature values do not match the evidence explanation, they will go with the feature values to make the final decision. Some examples that the study participants in the recommendation-driven condition go against the decision aid's prediction:

> "*Here, I believe the decision aid is mistaken. My rating would be medium because the house is very old which is overlooked by the model. Other features are all decent or above decent but the house age is an important feature.*" – Q11

> "*The location of the property is low so I thought that would bring down the price*" – Q7

Ignoring evidence is, of course, a good strategy if the decision maker believes that the evidence is wrong. However, ***recommendation-driven does not help participants to be aware of the high uncertainty among multiple predictions***. This limitation is mitigated by the hypothesis-driven paradigm.

For the AI explanation only paradigm, participants often rely on the feature values and not on the evidence explanation to make a decision. This is not surprising because participants can find it difficult to interpret the evidence without seeing the label that the evidence is referred to. We attribute this to the cognitive effort to link evidence to hypotheses, leading to participants ignoring evidence and relying on input feature values to make their decisions. This is a noteworthy limitation of *AI-explanation-only* as it makes people overlook the explanation if the link to the evidence is unclear. In the study by Gajos and Mamykina [15], the link from feature attributions to the task solution is more straightforward than in our study, which may explain the divergence of results.

Participants more often use the evidence to make a decision in the hypothesis-driven approach than in recommendation-driven or AI explanation only. This shows participants took advantage of the model evidence. In the two baseline conditions, participants tended to ignore evidence seemingly due to the inability to interpret it, which means they will fail to take advantage of the underlying model. In Figure 1, there are only two exceptions at Q6 and Q11 where the evidence is not the most used in the hypothesis-driven condition.

We found that in hypothesis-driven, participants reported that it was difficult to make decisions for two main reasons:

- ***Uncertainty awareness***: This is where there are multiple hypotheses with similar strength evidence. Participants are aware of the uncertainty in the model solely based on the positive and negative evidence provided for all hypotheses. In this case, participants use the input feature values or choose the hypothesis that they think is slightly better than the others when making the final decision. Figure 3 shows an example where two hypotheses *low* and *medium* both have a positive and negative weight of evidence, especially

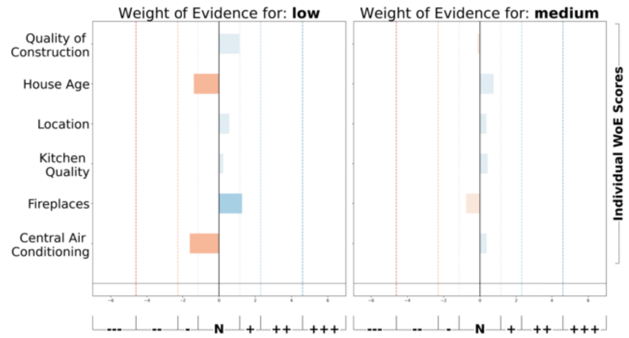

Figure 3. An example of ***uncertainty awareness*** (Q6).

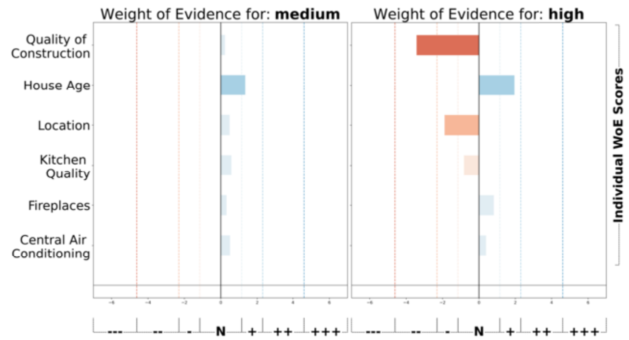

Figure 4. An example of ***deceptive evidence*** (Q9).

in the top three important features. For instance, some participants explicitly explain their uncertainty in the text as follows.

> "*I was choosing between high and medium. Quality of construction, age and location are the most important features. When it was at high, these were all positive. Kitchen quality and fireplaces were negative, but these are not as important.*" – Q0

> "*The amount of negative or positive evidence for low or medium is the same, including the three more important factors. Both medium or low could be viable but medium has less variance and is overall more balanced.*" – Q6

> "*The house is clearly not in a high bracket, but it is somewhat difficult to decide between low and medium. There are stronger indicators in low, going both ways, while medium has largely insignificant indicators. Low has a significant negative weighting for house age and this pushed me towards medium.*" – Q6

- ***Deceptive evidence***: When the evidence was strongest for an incorrect option. In this case, many participants just follow the evidence and make the wrong decision. Figure 4 illustrates an example of Q9 where we have all positive evidence in hypothesis *medium*, but strong negative evidence in hypothesis *high*. Therefore, all participants choose hypothesis *medium*, but hypothesis *high* is the ground truth. Future work will need to address the challenge of building trustworthy evidence.

In summary, the qualitative analysis showed that participants took advantage of the decision aid more in the hypothesis-driven condition than in recommendation-driven and explanation-only conditions. Further, we also found that participants recognised model uncertainty in the hypothesis-driven condition. However, there still remains a limitation of having deceptive evidence.

6 Discussions

Strengths and weaknesses of our hypothesis-driven approach First, participants using the hypothesis-driven approach required a similar time to complete the task compared to the recommendation-driven approach. Participants in the hypothesis-driven condition also made higher quality decisions than *recommendation-driven* and *AI explanation only* based on the Brier score. The results indicated that the *hypothesis-driven* gave study participants a more complete picture of the underlying decision aid than the other two approaches, helping them to make use of the AI models when they are right, and be less confident when the models are wrong.

Moreover, hypothesis-driven reduced over-reliance significantly compared to the standard AI recommendation. Similarly, hypothesis-driven also reduced under-reliance compared to AI explanation only. Importantly, the positive result for under-reliance using recommendation-driven is not cancelled out by the poor over-reliance result, compared to hypothesis-driven. The primary aim indicating potential for the use of uncertainty/confidence [5, 23] and conformal prediction [32] to direct decision makers' attention towards a set of hypotheses that it is confident about.

Using the qualitative analysis, hypothesis-driven helped participants take advantage of the decision support tool's evidence, and also recognise the uncertainty underlying the model. Using the strength of evidence, participants are aware of the uncertainty between multiple hypotheses. Therefore, they made an attempt to gauge the model uncertainty by calibrating the weight of evidence depending on whether the feature is important or not. Also, they could make use of the input feature values and choose the hypothesis that they perceive most likely matches with those values.

On the other hand, *recommendation-driven* and *AI explanation only* do not support this. We found that in recommendation-driven, people could use feature values to confirm the validity of the decision aid's prediction. However, they are not aware of the uncertainty among different hypotheses. In AI explanation only, people often ignore using the evidence and solely focus on using the feature values to make a decision because interpreting the evidence with this approach can be a lot more mentally demanding.

Study limitations There are also some limitations with the study. First, we ran the experiment on one dataset (Ames Housing), which limits generalisability. In addition, as there is no ground truth for the price of a house, the experimental participants' tasks are somewhat subjective. Further, this task has only three output classes, so only three hypotheses, and we anticipate the results would be more interesting when we consider more hypotheses. Finally, the human experiment is currently conducted with laypeople while experts likely interact with the decision-aiding tool differently from laypeople [14].

7 Conclusions

In this paper, we show that the hypothesis-driven approach using Weight of Evidence (WoE) can significantly reduce reliance, improve decision-making quality compared to two other prevalent decision-making approaches (recommendation-driven and AI explanation only). Furthermore, hypothesis-driven helps participants to be aware of the uncertainty among multiple options. Nevertheless, there still remains a challenge of study participants relying on the wrong (or misleading) evidence. Therefore, future work can address this challenge by exploring different approaches for presenting trustworthy evidence. More generally, potential future work is to consider the uncertainty/confidence in the generated evidence.

Acknowledgements

This research was supported by the University of Melbourne Research Scholarship (MRS) and partly funded by Australian Research Council Discovery Grant DP190103414. We also want to thank the anonymous reviewers for their valuable feedback.

References

[1] A. Abdul, C. von der Weth, M. Kankanhalli, and B. Y. Lim. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, 2020.

[2] D. Alvarez Melis, H. Kaur, H. Daumé III, H. Wallach, and J. Wortman Vaughan. From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9(1):35–47, 2021.

[3] D. Alvarez Melis, H. Kaur, H. Daumé III, H. Wallach, and J. Wortman Vaughan. User Study on Interpretability - Tutorial. https://github.com/dmelis/interpretwoe/blob/master/notebooks/WoE_UserStudy_Tutorial.ipynb, 2021. Accessed: 2023-05-30.

[4] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–16, 2021.

[5] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, M. Srikumar, A. Weller, and A. Xiang. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 401–413, 2021.

[6] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 2021.

[7] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.

[8] C. Chen, S. Feng, A. Sharma, and C. Tan. Machine Explanations and Human Understanding. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 2023.

[9] V. Chen, Q. V. Liao, J. Wortman Vaughan, and G. Bansal. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 2023.

[10] C.-W. Chiang and M. Yin. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 148–161, 2022.

[11] D. De Cock. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19 (3), 2011.

[12] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 2015.

[13] R. Fogliato, A. Chouldechova, and Z. Lipton. The Impact of Algorithmic Risk Assessments on Human Predictions and Its Analysis via Crowdsourcing Studies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 2021.

[14] R. Fogliato, S. Chappidi, M. Lungren, P. Fisher, D. Wilson, M. Fitzke, M. Parkinson, E. Horvitz, K. Inkpen, and B. Nushi. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1362–1374, 2022.

[15] K. Z. Gajos and L. Mamykina. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*, IUI '22, pages 794–806, 2022.

[16] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1), 2021.

[17] I. Good. Weight of evidence: A brief survey. In J. Bernardo, M. DeGroot, D. Lindley, and S. A.F, M, editors, *Bayesian statistics*, volume 2, pages 249–270. Elsevier, 1985.

[18] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*, 11(1), 2021.

[19] T. Kulesza, S. Stumpf, W.-K. Wong, M. M. Burnett, S. Perona, A. J. Ko, and I. Oberst. Why-Oriented End-User Debugging of Naive Bayes Text Classification. *ACM Trans. Interact. Intell. Syst.*, 1(1), 2011.

[20] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 126–137, 2015.

[21] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1369–1385, 2023.

[22] K. A. Lambe, G. O'Reilly, B. D. Kelly, and S. Curristan. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Quality & Safety*, 25(10):808–820, 2016. ISSN 2044-5415.

[23] T. Le, T. Miller, R. Singh, and L. Sonenberg. Explaining Model Confidence Using Counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):11856–11864, 2023.

[24] T. Le, T. Miller, L. Sonenberg, and R. Singh. Towards the New XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence, 2024. Full version of this paper. Available at arXiv:2402.01292.

[25] T. Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257, 2007.

[26] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, 2017.

[27] T. Miller. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 333–342, 2023.

[28] B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D. S. Wishart, A. Fyshe, B. Pearcy, C. MacDonell, and J. Anvik. Visual Explanation of Evidence in Additive Classifiers. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence - Volume 2*, pages 1822–1829, 2006.

[29] C. Rastogi, Y. Zhang, D. Wei, K. R. Varshney, A. Dhurandhar, and R. Tomsett. Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), 2022.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[31] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, 6th edition, 2016.

[32] E. Straitouri, L. Wang, N. Okati, and M. G. Rodriguez. Improving expert predictions with conformal prediction. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pages 32633–32653, 2023.

[33] A. T. Tan. Cracking the Ames Housing Dataset with Linear Regression. https://github.com/at-tan/Cracking_Ames_Housing_OLS", note="Accessed: 2023-05-30, 2021.

[34] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[35] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, 2021.

[36] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

[37] M. Vered, T. Livni, P. D. L. Howe, T. Miller, and L. Sonenberg. The Effects of Explanations on Automation Bias. *Artificial Intelligence*, 2023.

[38] X. Wang and M. Yin. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Trans. Interact. Intell. Syst.*, 2022.