February 2024

Measuring the Effect of Mental Workload and Explanations on Appropriate AI Reliance Using EEG

Zelun Tony ZHANG [a,1], Seniha KETENCI ARGIN [b], Mustafa Baha BILEN [b], Doğan URGUN [b], Sencer Melih DENIZ [b], Yuanting LIU [c] and Mariam HASSIB [c]

[a] *fortiss GmbH, Research Institute of the Free State of Bavaria and LMU Munich*
[b] *The Scientific and Technological Research Council of Turkey (TÜBITAK) Informatics and Information Security Research Center (BILGEM)*
[c] *fortiss GmbH, Research Institute of the Free State of Bavaria*

Abstract. AI is anticipated to improve human decision-making across various domains, often in high-stakes, difficult tasks. However, human reliance on AI recommendations is often inappropriate. A common approach to address this is to provide explanations about the AI output to decision makers, but results have been mixed so far. It often remains unclear when people can rely appropriately on AI and when explanations can help. In this work, we conducted a lab experiment ($N = 34$) to investigate how the appropriateness of human reliance on (explainable) AI depends on the mental workload induced by different decision difficulties. Instead of self-assessments, we used EEG (Emotiv Epoc Flex head cap, 32 wet electrodes) to more directly measure participants' mental workload. We found that the difficulty of a decision, indicated by the induced mental workload, strongly influences participants' ability to rely appropriately on AI, as assessed through relative self-reliance, relative AI reliance, and decision accuracy with and without AI. While reliance was appropriate for low mental workload decisions, participants were prone to overreliance in high mental workload decisions. Explanations had no significant effect in either case. Our results imply that alternatives to the common "recommend-and-explain" approach should be explored to assist human decision-making in challenging tasks.

Keywords. human-AI decision-making, explainable AI, appropriate reliance, decision difficulty, mental workload, EEG

1. Introduction

AI is expected to enhance human decision-making in a wide variety of domains, including high-stakes ones like law enforcement [1] or healthcare [2]. However, people's reliance on AI-generated decision recommendations is often inappropriate, with multiple studies observing either *underreliance* (rejection of correct/beneficial recommendations) [3,4,5] or *overreliance* (acceptance of incorrect/detrimental recommendations) [6,7,8]. Such inappropriate reliance prevents humans and AI from complementing

[1] Corresponding Author: Zelun Tony Zhang, zhang@fortiss.org.

each other [9], and in the case of high-stakes decisions also raises ethical concerns. A common approach is to try to enable *appropriate reliance* by explaining the AI recommendations. The rationale is that by making transparent how the AI produced a recommendation, people can better judge when the recommendation is reliable or not. However, results have been mixed so far. While a few studies do find more appropriate reliance with explanations [10,11], in the majority of studies, explanations were either ineffective [12,13,14] or even increased overreliance [9,15,16,17]. Often, it remains unclear *when* people are able to rely appropriately on AI and when explanations can improve the appropriateness of reliance. One possible factor could be the difficulty of a decision, as overreliance and the ineffectiveness of explanations to mitigate it have been particularly observed in tasks that are challenging for humans [12,13,16,17].

In this work, we aimed to investigate how the mental workload of solving a decision task affects appropriate reliance and the effectiveness of explanations. Such an understanding would be of interest as it is often hoped that AI can improve human decision-making in difficult decisions that induce more mental workload, such as interpreting medical images [18,19,20]. The call for explainable AI is particularly prominent in such critical applications, even though its effectiveness is unclear [21]. We contribute to the discourse by posing the following research questions:

RQ1: How does the appropriateness of human reliance on AI depend on the mental workload induced by the difficulty of a decision task?
RQ2: How does the effect of explanations on the appropriateness of reliance depend on the mental workload induced by the difficulty of a decision task?

To answer our research questions, we conducted an experiment with the recognition of noisy images of varying difficulty as test bed to assess the appropriateness of participants' AI reliance, both with and without explanations. As a methodological contribution, we used electroencephalography (EEG) to measure participants' mental workload while looking at the images as indicator for their subjectively experienced decision difficulty, rather than asking for it after each decision. We found that for decisions with low mental workload (indicating easier decisions), people relied appropriately on AI, leading to *complementary team performance* (i.e., the performance of human and AI together is better than that of either alone). For decisions with high mental workload (indicating more difficult decisions), overreliance was much more pronounced, and team performance was not complementary. Explanations had no significant effect in our study, neither for low nor for high mental workload. Our results highlight the challenges of supporting human decision-making with AI in difficult, high-stakes decisions.

2. Related Work

2.1. Reliance on (Explainable) AI and Decision Difficulty

There are several studies related to how the difficulty of a decision influences people's reliance on AI. For instance, Lu and Yin [22] investigated how people derive heuristics for how to rely on AI based on decisions where they are highly confident and how they observe the AI to perform on those decision tasks. Papenmeier et al. [23] showed that AI errors on difficult, more ambiguous decisions hurt people's perception of the AI's ac-

curacy less than errors on easy decisions. Importantly to our study, Parkes [24] demonstrated that it is subjectively perceived task difficulty, not objective task complexity, that is positively related to reliance on decision aids. All these studies show that the difficulty of a decision task plays an important role in how humans assess the reliability of AI and in how they choose to rely on it. However, these studies do not directly assess the appropriateness of people's reliance behavior, and how the effect of explanations on reliance depends on difficulty.

We are aware of two closely related works that do address task difficulty in relation to appropriate reliance and explanations. Wang and Yin [12] compared the effectiveness of several explanation styles regarding appropriate reliance, among other desiderata, on two different decision tasks, one where participants had more domain knowledge and another where they had less. One can assume that the latter task was more challenging for participants. Vasconcelos et al. [10] explicitly studied how explanations and task difficulty interact to influence overreliance. Their study was based on a maze solving task, and task difficulty was manipulated through the complexity of the mazes, such that more difficult tasks required significantly more effort from participants.

In contrast to these two works, our goal was to vary the difficulty of decisions while keeping the task domain and complexity constant. This is more in line with the notion of difficulty studied by Lu and Yin [22] or Papenmeier et al. [23], where difficulty is not caused by a lack of domain knowledge or the required effort, but rather by the presence of more than a single plausible answer. We were more interested in this notion of difficulty as it better reflects the difficulty of many real-world decisions such as medical diagnoses or creditworthiness assessments. To clarify that we study difficulty due to the presence of more than one plausible answer, we use the term *decision difficulty* throughout this paper.

We further did not treat decision difficulty as an objective property of a decision task, since the same decision might be challenging for one person, but not for another. Instead, we chose to account for this subjectivity by using EEG to measure participants' mental workload while making a decision, which we used as indicator for subjective decision difficulty. To the best of our knowledge, this is the first work investigating how the appropriateness of human reliance on AI and the effectiveness of explanations to improve appropriate reliance is affected by subjective, decision difficulty-induced mental workload. The use of physiological measures further sets our work apart from related studies like the one by Lu and Yin [22], who asked participants to rate their confidence after each decision. Compared to such a self-rating approach, our EEG-based method has the advantage that it does not introduce a secondary task (rating the difficulty of a decision), which may distract from the primary task (making the decision).

2.2. Mental Workload Measurement With EEG

Mental workload (MWL) signifies the level of engagement of a limited set of cognitive resources during the ongoing processing of a primary task, influenced by various external factors such as environmental conditions and situational variables, as well as by intrinsic traits of the human operator. It involves the allocation of effort and attention to manage the constant demands of the task [25]. In addition to self-reported measures and primary task performance measures, physiological measures are often applied in assessing MWL. Physiological measures assess MWL through the analysis of physiological responses of

an operator while executing a primary task [26,27,28,29,30,31]. Among physiological measures, EEG is widely utilized in assessing MWL due to its direct measurement of signals from the brain, as opposed to the indirect measurement of other physiological responses initiated by the brain [28,32,33].

Recent research has prioritized validating the utilization of physiological responses to quantify individuals' MWL [34,35]. Evidence indicates that neurophysiological measurements, such as EEG signals, exhibit a direct correlation with the mental demand encountered during tasks [36]. The electrical activity within the prefrontal cortex, as measured by the theta frequency band, escalates in tandem with increasing cognitive demands [37], while activity in the parietal midline alpha band decreases [34,38,39]. EEG spectral components consistently display discernible variations in response to varying cognitive task requirements [40,41,42], underscoring a correlation between EEG spectral power and task intricacy.

Significant progress has been made in understanding MWL, with key contributions from Sweller's Cognitive Load Theory and subsequent developments by Paas, Renkl, and Sweller [43,44]. Relevant to our methodology (Section 3), recent research has explored how mental workload is affected by visual stimuli degradation [45,46] and how mental workload can be measured from EEG signals during simulated tasks. A notable study by Kartali et al. showed a strong correlation between EEG-derived mental workload metrics and task complexity, providing valuable insights for real-time monitoring of cognitive processing in complex tasks [47].

3. Method

Our aim was to investigate the effect of explanations and the mental workload induced by varying decision difficulties on human reliance on AI. To this end, we conducted an experiment in which participants had to solve a series of carefully selected decision tasks with the help of AI predictions, both with and without explanations. We collected EEG data along with participants' decisions in conjunction with the AI predictions to derive the mental workload induced by different decision difficulties and participants' reliance behavior. Figure 1 gives an overview of our study and how the rest of the paper maps onto it.

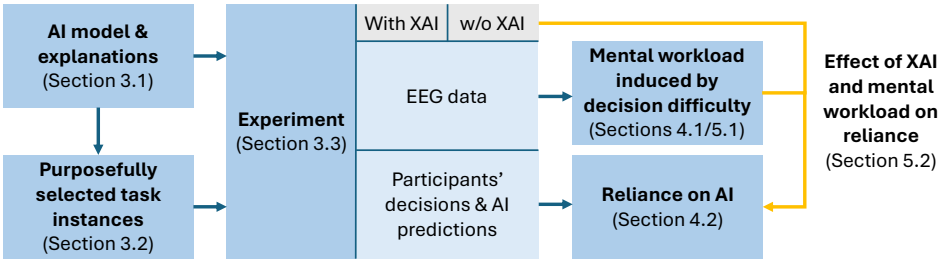


Figure 1. Flow chart of the study and the respective sections of the paper.

In this section, we describe our methodology. We start by presenting the decision-making task (Section 3.1) and our procedure for selecting the specific task instances for the experiment (Section 3.2). We close the section by presenting the experimental design and procedure (Section 3.3).

3.1. Human-AI Decision-Making Task

To investigate our research questions, we chose an image recognition task where participants had to decide with the help of AI predictions what kind of objects are shown in a range of noisy images. We utilized the image dataset presented by Steyvers et al. [48] as the basis for our image classification task. The dataset is composed of 1200 unique images over 16 categories (e.g., boat, dog, airplane, etc.) that are a subset of the ImageNet Large Scale Visual Recognition (ILSRVR) dataset [49]. To create varying difficulties, Steyvers et al. applied four different levels of phase noise to each image. In addition to the ground truth labels, Steyvers et al. conducted an experiment to measure user accuracy in classifying the images. They collected confidence levels (low, medium, high), task accuracy, and task completion time, from 145 Amazon Mechanical Turk workers [48]. Comparable image classification accuracy was achieved by both humans and AI [48]. This image recognition task has been used in prior studies investigating human-AI decision making [50,51]. We chose this task as it is a generic task that does not require participant expertise in a certain domain, hence increasing the generalizability of our experimental results. The task was also suitable for our purpose, as visual stimuli affect cognitive workload in a way that could be measured by EEG [45].

To generate AI predictions that would be later presented to participants in our experiment, we used a DenseNet-161 model [52] which was pre-trained on ImageNet and fine-tuned on our dataset for 100 epochs, following the approach by Hemmer et al. [50]. We used a training-validation-test split of 60%, 20%, and 20% respectively, and achieved a test set accuracy of 0.913. The model was implemented using PyTorch 2.0.1 [53] and trained on a virtual machine in the free version of Google Colab[2] with 12 GB RAM and a Tesla T4 GPU with 16 GB memory. For explanations, we displayed saliency areas in the form of heatmaps superimposed on the images, which is one of the most common ways to explain image classification models. We generated the heatmaps using the popular GradCAM algorithm [54], implemented with the PyTorch GradCAM library [55].

3.2. Task Instances Selection

As our goal was to study how reliance on AI depends on the mental workload induced by varying decision difficulties, we needed to ensure that the task instances presented to our participants covered a wide range of decision difficulties. Hence we combined different criteria available in Steyvers et al.'s dataset to assign difficulty levels to each image. Based on these pre-assigned levels, we would choose a representative sample of varying decision difficulties that would induce different mental workload. We used the following features from Steyvers et al's dataset [48]:

- Image phase noise level (80, 100, 110, 125)
- Crowd workers' confidence (low, medium, high)
- Crowd workers' decision time
- Crowd workers' accuracy
- Crowd workers' agreement, which we derived by applying Fleiss' Kappa [56] to crowd workers' individual answers present in the dataset.

[2]https://colab.research.google.com/

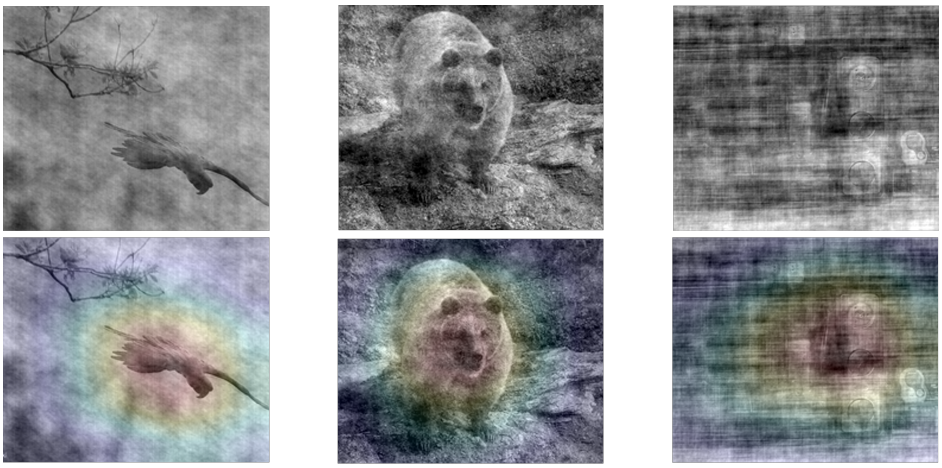


Figure 2. Example images from our dataset: The top row shows three images, without explanations, the bottom row shows the images with the GradCAM explanations. All three images were classified correctly using our model.

We then applied principal component analysis (PCA) to reduce the features to two dimensions and used k-means clustering to create three clusters: easy, medium, and hard.

For our final set of task instances, we then randomly selected 240 images, 80 images for each of the three difficulties, where 64 were classified correctly by the AI model and 16 were classified incorrectly, such that participants experienced an accuracy of 80%. For the purpose of our study, we reduced the number of class choices per image that would be presented to participants to the top four choices predicted by the AI model. If the top four choices did not contain the correct ground truth label, we replaced the fourth choice with the correct label.

Figure 2 shows examples of correctly classified easy/medium/hard images and their GradCAM explanations. We reiterate that these pre-assigned difficulties were only meant to ensure that our experiment covered decisions of varying difficulties. How difficult a decision is actually perceived by a participant is subjective to an extent and may deviate from the pre-assigned difficulties. This is why we used EEG to measure mental workload as indicator for the subjectively experienced decision difficulty, as explained in the following sections.

3.3. Experimental Design and Procedure

In this section we outline our experimental design and explain our procedure in detail. We designed a within-subjects experiment with two decision-making conditions: *with explainable AI* and *without explainable AI*, which we will refer to as *XAI* and *NXAI* for the rest of this manuscript. The procedure was divided into two main sessions (XAI, NXAI) which were counterbalanced to avoid sequence effects. Therefore, half of the participants started with the XAI session, while the remaining half started with the NXAI session. Within each session, three sets of 40 tasks were presented, one set for each difficulty level. The order of the sets of tasks, the order of the images in each set, and the

assignment of an image to the XAI or NXAI session were also counterbalanced. Figure 3 depicts the experimental design in detail.

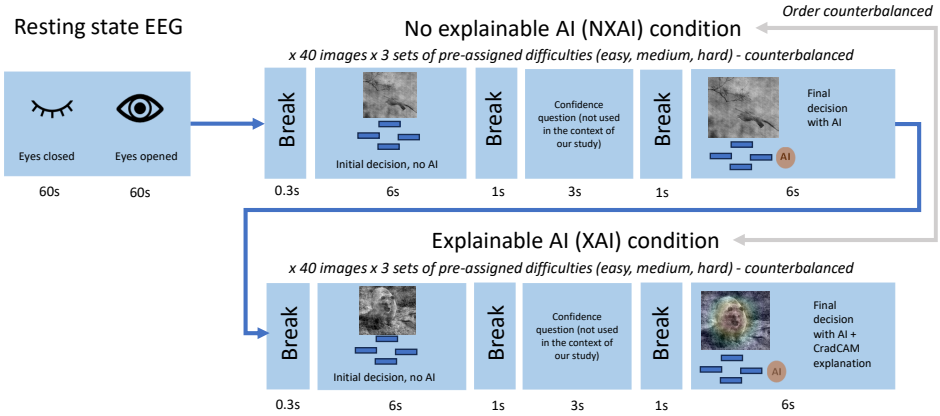


Figure 3. Schematic overview of the experimental design.

After arriving at our lab, participants were informed about the data acquisition procedure, gave written consent before the experiment and received monetary compensation for their participation. This study complied with the ethical standards outlined in the Declaration of Helsinki and was approved by the Scientific Research and Publication Ethics Board for Science and Engineering of Karabük University (petition number: 346383). To collect EEG data, we used the Emotiv Epoc Flex EEG head cap[3] with 32 wet active electrodes (Fp1, Fp2, F7, F8, F3, F4, Cz, Fz, FC1, C3, FC5, FT9, T7, CP5, CP1, P3, P7, PO9, Pz, PO10, P8, P4, CP2, CP6, T8, FT10, FC6, C4, FC2, Oz, O1, O2) positioned according to the 10–20 international system. Two additional electrodes were used for reference in the ears and conductive gel was applied to enhance electrode-to-scalp contact. The sensors are made from sintered silver/silverchloride to minimize the electrode impedance and are fully compatible with gel electrolytes. During the study, participants were seated in front of a monitor (HP E233, 1920×1080 resolution) at approximately 50 cm distance and instructed to focus on the center of the screen, as shown in Figure 4. The first session started with an eyes-closed eyes-opened task, each for 60 seconds, typical for EEG studies, for collecting resting EEG data.

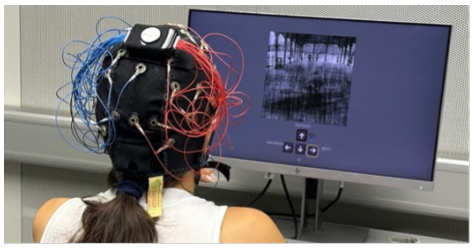
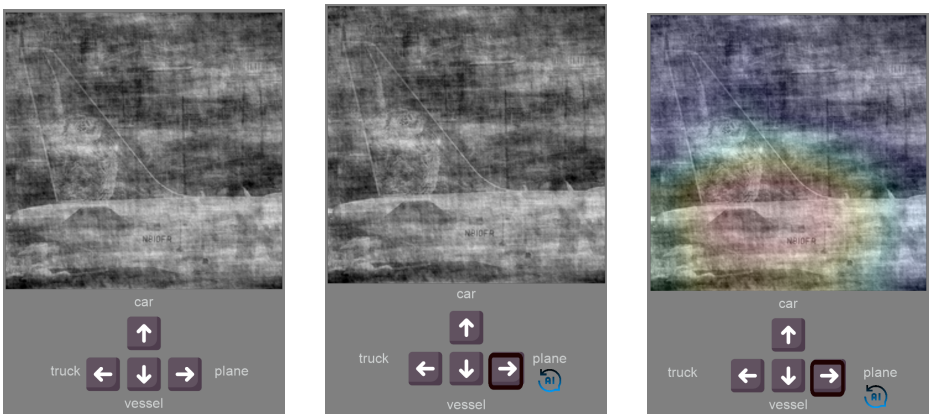


Figure 4. EEG data acquisition environment.

[3]Emotiv Epoc Flex: `https://www.emotiv.com/flex-gel`

The screen first showed a visual (fixation cross) and auditory (beep sound) cue for 0.3 seconds. After the fixation cross disappeared, a new task stimulus was introduced. Participants had to first make an initial decision without AI recommendations. They were presented with an image from our dataset and shown four choices as to what the image depicted (Figure 5a). For the purpose of EEG data acquisition, participants had a fixed time of six seconds to choose an answer. We tested this time interval in pilot tests and found it to be sufficient to respond to the task. Following their initial response, participants answered a question about the confidence of their decision, with two levels of confidence available for selection. The confidence level question is not part of the focus of this paper and hence we do not report on its results later. Subsequently, the same image was presented again, this time accompanied by the AI model's classification for that image, presented by highlighting the respective option and showing an "AI" icon next to it. In the XAI condition, the AI recommendation was complemented by the GradCAM explanation overlayed on top of the image (Figure 5c). In the NXAI condition, the AI recommendation was shown with the original image again (Figure 5b).



(a) Initial decision without AI.

(b) Final decision with AI in the NXAI condition.

(c) Final decision with AI in the XAI condition.

Figure 5. Screenshots of task interface used during the study. Participants chose an answer using the respective arrow key on the keyboard. Choices were originally presented in participants' native language and translated for this figure.

4. Measures

In this section, we introduce our EEG analysis pipeline and mental workload measurement approach (Section 4.1). We further explain the reliance metrics (Section 4.2) which are later used to present our outcomes in the results section.

4.1. EEG Measures of Mental Workload

Figure 6 depicts the EEG analysis pipeline used for processing EEG data and measuring mental workload. We first collected EEG data sampled at 128 Hz, as explained in Sec-

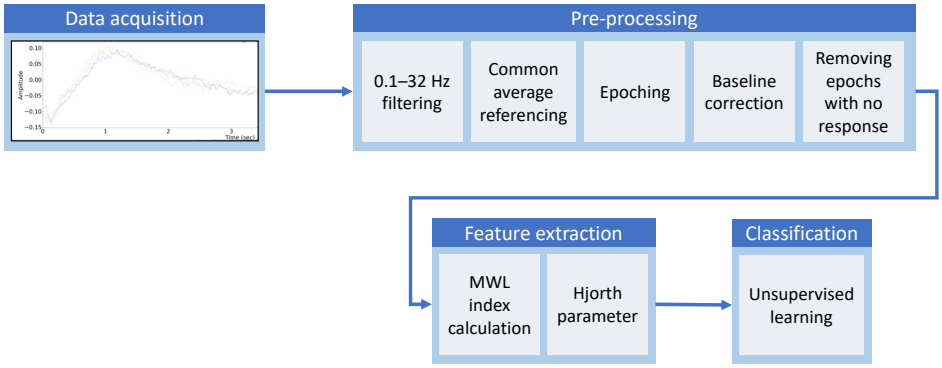


Figure 6. EEG signal analysis steps.

tion 3.3. In the data cleaning and pre-processing phase, we removed participants with more than ten incomplete responses. To extract the relevant frequency band information, we band-pass filtered the EEG data from 0.1 Hz to 32 Hz. In this way, we also suppressed line noise. To remove eye blink and facial movement artifacts, we applied common average referencing (CAR), a re-referencing process [57] that can increase the signal-to-noise ratio [58]. In short, it is obtained for each channel by subtracting the average of all other channels from the current one. Compared to other methods like independent component analysis (ICA), which can be inconsistent in improving EEG data quality [59] and can distort the EEG signal [60], CAR better preserves the precise channel-specific information needed for our purpose [61,62,63,64]. Epochs were created from 0.3 seconds before to 6 seconds after stimulus onset, encompassing the entire EEG response to the stimulus, with each epoch aligned to the stimulus presentation. Finally, we performed baseline correction for each epoch, from 200 ms before the stimulus to the onset of the stimulus, to eliminate any DC offset.

In the feature extraction phase, we utilized the power of different EEG frequency bands. When the complexity of tasks rises, there is an observable augmentation in the frontal midline theta band (4–7 Hz) and a concomitant reduction in the parietal midline alpha band (8–12 Hz) [34,38,39]. With respect to the related frontal and parietal areas, we chose the F_z and P_z channels for the measurement. To obtain theta power of F_z and alpha power values of P_z for each 6-seconds EEG trial, which were to be used in the calculation of the MWL index, we decomposed the EEG trails into frequency components for each epoch using the fast fourier transform (FFT). A power spectral density (PSD) was obtained for each EEG trial using a box window of 3 seconds in length with a 2.9 seconds overlap. Using this overlap interval, we obtained 58 windows for each EEG trial. The relative band powers for the theta and alpha frequency bands for each window were calculated using the individual alpha frequency (IAF) method [65]. The MWL index was then computed by dividing the relative power of the alpha band in the P_z channel by the relative power of the theta band in the F_z channel, as detailed in Equation 1, offering a measure of cognitive effort and engagement during the task [47,66,67,68].

$$MWL = \frac{\theta_{F_z}}{\alpha_{P_z}} \tag{1}$$

Based on the 58 MWL index values per EEG trial, we obtained a feature set for each EEG trial. We calculated the root mean square (RMS) and Hjorth activity parameter [69] of the 58 MWL index values and used those as features for the related EEG trial. Hjorth parameters are statistical measures that enable the analysis of signals in the time domain [70]. These parameters include activity, mobility, and complexity. We utilized the activity parameter, as it indicates the signal power, defined as the variance of a time function. It reflects changes in the power spectrum surface within the frequency domain. To clarify, the activity parameter takes a large value if the signals contain many high-frequency components, and a small value if they consist of few. The activity parameter is calculated using Equation 2, where $x(n)$ is the input signal and $\underline{x}$ is the average of the signal.

$$Activity = \frac{1}{N}\sum_{n=0}^{N-1}[x(n)-\underline{x}]^2 = var(x) \tag{2}$$

In the classification phase, we aimed to categorize the mental workload for all epochs of each participant separately. Most current approaches for classifying tasks based on EEG signals require extensive labeled training datasets. Yet, gathering and manually labeling a vast array of EEG recordings from a large number of participants is impractical. To mitigate this issue, we employed unsupervised learning to classify cognitive workload in the EEG recordings (see Section 5.1). This approach is informed by established theories in cognitive load and decision-making [71,72,73,74,75,76,77,78,79], and also addresses the potential for subjective bias in manual labeling, highlighting the benefits of unsupervised learning in the context of EEG analysis [80,81].

4.2. AI Reliance Measures

We used the measurement concept proposed by Schemmer et al. [14] to assess the appropriateness of reliance on AI. The measurement concept requires a setup where the user first makes an initial decision without AI, and then receives the AI recommendation before making the final decision. Based on users' initial and final decisions as well as the correctness of the AI recommendation, two dimensions are defined to assess the appropriateness of users' AI reliance: *relative self-reliance* (RSR) and *relative AI reliance* (RAIR). The former measures the degree to which users either correctly rejected wrong AI recommendations (high RSR) or overrelied (low RSR), while the latter measures the degree to which they either adopted correct AI recommendations (high RAIR) or underrelied (low RAIR).

Schemmer et al. considered binary decisions, so we adapted their measurement concept according to Figure 7 to accommodate cases where human and AI chose different answers, but both are wrong; or cases where the human revises their answer after seeing the AI recommendation, but both the intial and final answer are wrong. We further adapted Schemmer et al.'s definitions of RSR and RAIR to accommodate our adaptations as follows:

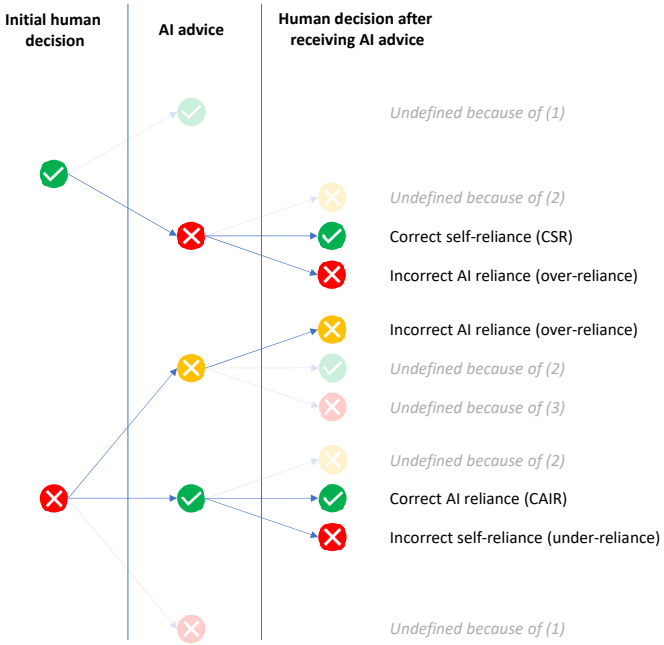


Figure 7. Possible combinations of human decisions and AI advice in our study. In line with the argumentation of Schemmer et al. [14], we exclude cases where (1) the initial human decision is the same as the AI advice, or (2) where the final human decisions is different from both the initial decision and the AI advice. We further exclude cases where participants held on to their initial wrong decision after getting AI advice that was wrong, but different from their initial decision (3). These cases are ill-defined under Schemmer et al.'s framework, as participants incorrectly self-relied, but did *not* under-rely, since the AI was also wrong. The remaining cases mostly correspond to the definitions in [14], with the exception of cases where participants adopted wrong AI advice after both the initial human decision and the AI advice were wrong, but different. We count this as an additional path for overreliance. Figure adapted from [14].

$$RSR = \frac{\sum_{i=0}^{N} CSR_i}{\sum_{i=0}^{N} CSR_i + \sum_{i=0}^{N} OR_i} \tag{3}$$

$$RAIR = \frac{\sum_{i=0}^{N} CAIR_i}{\sum_{i=0}^{N} CAIR_i + \sum_{i=0}^{N} UR_i} \tag{4}$$

Schemmer et al. defined RSR with those cases in the denominator where the initial human decision was correct, but the AI recommendation was wrong. They correspondingly defined RAIR with those cases in the denominator where the initial human decision was wrong and the AI recommendation was correct. For the binary decision case, these are equivalent to the denominators in Equations 3 and 4. For our case, the adaptation of the denominator in Equation 3 factors in the additional path for overreliance in Figure 7. The adaptation in Equation 4 has no consequences, but we made it to keep the definitions consistent.

5. Results

We recruited 45 healthy participants for the study. After removing data from participants with more than ten incomplete responses, we retained the data from 34 participants (6 females, 28 males) with an average age of 25.9 ($SD = 1.4$) years; four participants were left-handed.

5.1. Analysis of EEG Data and Mental Workload

With the method described in Section 4.1, we were able to assess participants' mental workload during both their initial decision without AI and the final decision with AI support. We argue that the former is a more faithful measure of the mental workload induced by the task itself and therefore focus our analyses hereafter on the mental workload measures obtained during the initial decision phase.

In an initial analysis step, we calculated the MWL indices for each epoch as described in Section 4.1 and averaged the MWL values across easy, medium, and hard images respectively for each participant. Figure 8 exemplarily shows the resulting data for Participants 1 and 2, showing that easy and hard images induce MWL index values that are clearly separable, while the MWL from medium images is much less distinguishable.

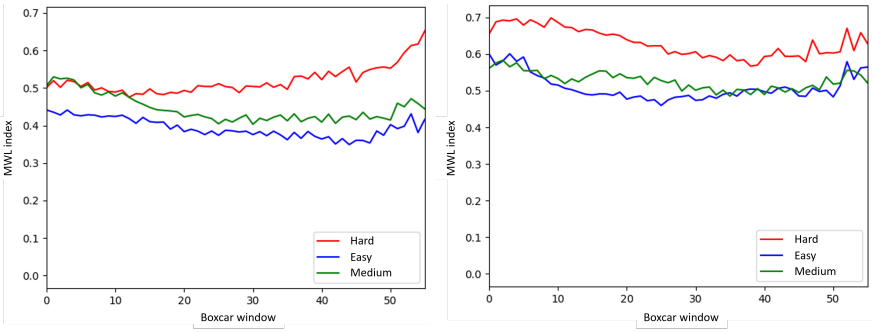


Figure 8. Mental workload index values for Participant 1 (left) and Participant 2 (right).

Following this initial observation, we applied binary, unsupervised classification to the EEG data to distinguish between low and high mental workload. We explored a range of common unsupervised classification algorithms, including k-means [71,72,76,78,79], mean shift [74], Gaussian mixture [77], agglomerative [75], and spectral clustering [73], to analyze how the data clustered according to MWL index values and the Hjorth activity parameter for each epoch [69]. Figure 9 exemplarily shows the resulting clusters for Participant 1, and Table 1 shows the share of images where the clustering into low/high mental workload corresponds to the easy/hard pre-assigned difficulties from Section 3.2, averaged over all participants. Note that images of medium pre-assigned difficulty were not considered for Figure 9 and Table 1 due to their unclear mapping to the mental workload clusters, but *were* included in the reliance analyses in Section 5.2. The high correspondence between easy/hard pre-assigned decision difficulties and the EEG-based low/high MWL clusters validates that our unsupervised classification approach is capable of identifying differences in mental workload induced by varying decision difficulties.

For the further analyses in Section 5.2, we chose the k-means clustering results due to their high correspondence with the pre-assigned decision difficulties.

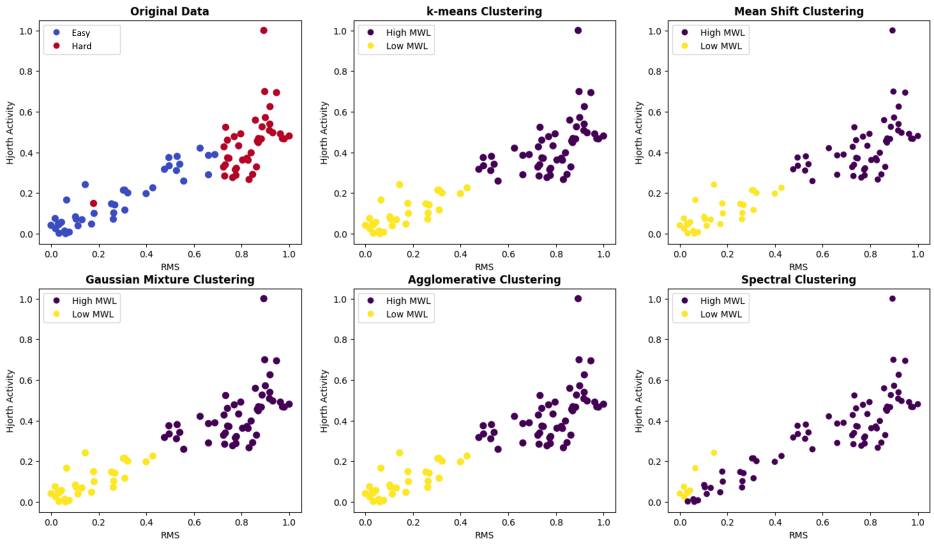


Figure 9. Clustering results of the different clustering algorithms for Participant 1.

Table 1. Correspondence between easy/difficult images and low/high mental workload, as classified by the different clustering algorithms.

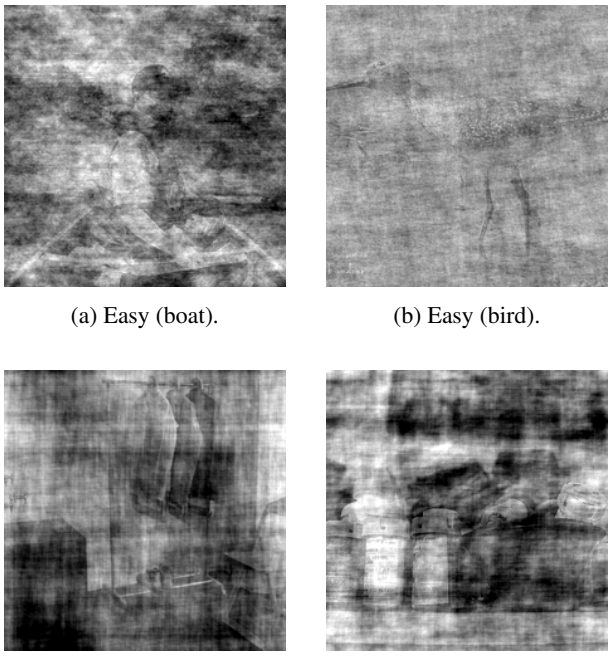
k-means clustering	Mean shift clustering	Gaussian mixture clustering	Agglomerative clustering	Spectral clustering
0.79	0.75	0.78	0.78	0.72

While a high correspondence between MWL clusters and the pre-assigned decision difficulties is encouraging, some amount of deviation is to be expected due the subjective nature of decision difficulty. One person might easily recognize what is shown in an image, while a different person might be unable to classify that same image. Figure 10 shows some example images where mismatches between the pre-assigned decision difficulty and the MWL classification were particularly frequent. This demonstrates the importance of considering the subject-dependent mental workload instead of pre-assigned decision difficulties in the subsequent analysis of participants' reliance behavior.

5.2. Analysis of AI Reliance

5.2.1. Effects of Explanations and Mental Workload on Reliance

As proposed by Schemmer et al. [14], we analyzed the appropriateness of reliance in terms of RSR and RAIR as two separate dimensions, as shown in Figure 11. We fitted mixed-effects logistic regression models with random intercepts and slopes for individual participants (see Appendix A) to investigate the effects of mental workload and study conditions on both metrics. For mental workload, we used the outcome from the unsupervised clustering into high and low mental workload decisions as measured by EEG



(a) Easy (boat). (b) Easy (bird).

(c) Hard (chair). (d) Hard (bottle).

Figure 10. Images with frequent mismatches between easy/hard pre-assigned decision difficulties, calculated as described in Section 3.2, and low/high EEG-based mental workload classifications. Ground truth classes in parentheses.

during participants' initial decision, as laid out in Section 5.1. We performed likelihood ratio tests on the regression models for statistical significance testing, and report odds ratios (referred to as *OR*) as effect sizes as well as their confidence intervals (referred to as *CI*).

Our results show that RSR is significantly lower for high mental workload decisions (OR = 0.101, 95% CI [0.026, 0.210], $\chi^2(1) = 33.167$, $p = 8.46 \times 10^{-9}$) indicating higher overreliance on AI. There was no statistically significant main effect of the study condition (*XAI*, *NXAI*) (OR = 1.029, 95% CI [0.640, 1.739], $\chi^2(1) = 0.013$, $p = 0.910$). Interaction effects were also not significant (OR = 0.491, 95% CI [0.162, 1.372], $\chi^2(1) = 1.952$, $p = 0.162$).

Considering RAIR, the results show no significant main effects for mental workload (OR = 0.787, 95% CI [0.547, 1.103], $\chi^2(1) = 1.494$, $p = 0.222$) , or study condition (OR = 1.029, 95% CI [0.640, 1.739], $\chi^2(1) = 0.013$, $p = 0.910$). Additionally, no interaction effects were found (OR = 0.908, 95% CI [0.485, 1.688], $\chi^2(1) = 0.086$, $p = 0.769$).

5.2.2. Complementary Team Performance

We further analyzed the outcomes of the reliance behavior described above in terms of decision accuracy. As shown in Figure 12, the AI accuracy was around 0.8 for both high and low mental workload decisions. Human accuracy was slightly higher, but comparable for low mental workload decisions. The final accuracy of humans supported by AI was higher than both humans and AI on their own, meaning complementary performance

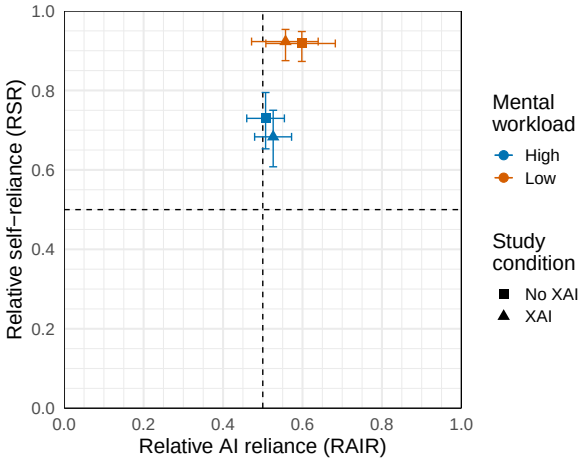


Figure 11. Appropriatenes of reliance. Lower RSR means more overreliance, while lower RAIR means more underreliance. Perfectly appropriate reliance would be in the upper right corner. Error bars denote 95% confidence intervals.

was achieved. On high mental workload decisions, humans were much less accurate than the AI, and complementary performance was *not* achieved, as human accuracy with AI support was lower than the AI's individual accuracy.

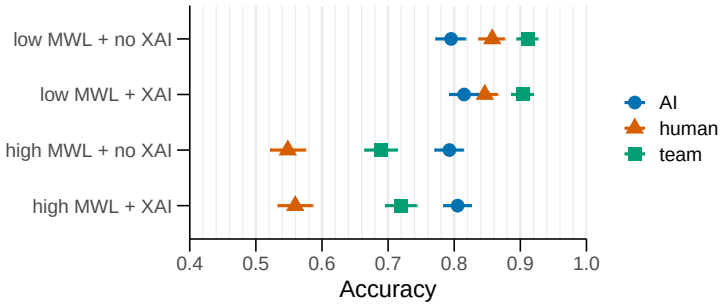


Figure 12. Decision accuracy of humans and AI individually and together. Human accuracy is derived from participants' initial independent decisions, team performance from participants' final decisions after seeing AI recommendations. Error bars denote 95% confidence intervals.

We fitted mixed-effects logistic regression models with random intercepts (see Appendix A) to investigate the effect of explanations on participants' AI-assisted accuracy, but likelihood ratio tests revealed no significant effect, neither for high mental workload decisions (OR = 1.143, 95% CI [0.965, 1.355], $p = 0.122$), nor for low mental workload ones (OR = 0.945, 95% CI [0.719, 1.242], $p = 0.689$).

6. Discussion

6.1. Appropriateness of Reliance Depends Strongly on Decision Difficulty

Our results suggest that human ability to rely appropriately on AI recommendations depends strongly on the difficulty of the decision task. For less challenging decisions, AI can provide a helpful second opinion that can complement human decision-making even without explanations. For such low mental workload decisions, RSR was high and RAIR comparatively low. This indicates that participants mostly remained with their initial decision as they could easily form an opinion about the correct answer, but occasionally noticed when it would be beneficial to switch to the AI recommendation.

For more challenging decisions, reliance was less appropriate, primarily driven by a strong increase in overreliance compared to low mental workload decisions, as indicated by the much lower RSR. Interestingly, the considerably higher overreliance was not accompanied by more *correct* AI reliance, meaning that participants did not adopt a "blind trust" policy, as is sometimes discussed in previous work [6,9]. Apparently, when faced with challenging decisions where people may have little clue about the correct answer, participants were more willing to switch to the AI recommendation, but also much more erratic in judging when to do so.

6.2. The Role of Explanations

While explanations had no significant effect in our experiment, the results still provide some insights regarding explanations. Saliency-based explanations like in our study are conceptually mostly unhelpful for detecting when the AI is wrong [82,83], as the saliency area often makes sense even when the model is wrong. However, intuitively, one could expect that the heatmaps would help users to recognize more instances when it is helpful to switch to the AI (see Figure 13), i.e., increase RAIR. This did not turn out to be the case, likely because such instances were too rare.

Still, these properties of saliency maps point to their potential to further improve the complementarity between human and AI for low mental workload decisions, since the deficit there was that participants often did not recognize when it was beneficial to adopt the AI recommendation. Recognizing AI error was less important since overreliance was already low. At the same time, the results show that saliency maps are conceptually not suitable for high workload decisions, since overreliance, i.e., failure to recognize when the AI was wrong, was the primary issue there, which is exactly where saliency-based explanations do not help.

6.3. Implications for AI-Assisted Decision-Making

Our results imply that for less challenging decisions, simply showing AI-generated recommendations may be sufficient to achieve complementary team performance between humans and AI. Explanations that effectively highlight when the AI is correct may help to further improve complementarity. This conforms to the common notion to increase trust in AI through explanations [84,85]. However, the aim in AI-assisted decision-making is often to support decisions that are difficult for humans. For these challenging decisions, our results imply that humans are significantly less able to rely appropriately on AI and are more likely to be misled by AI when it is wrong. Our results further imply that it

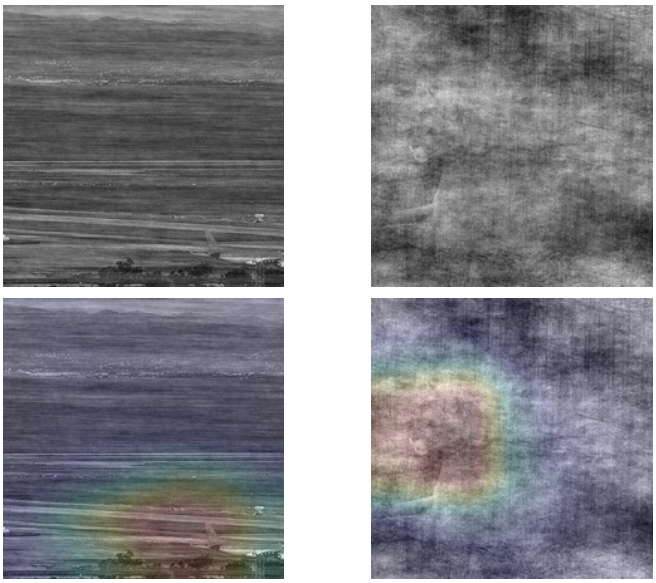


Figure 13. Examples where the heatmap explanation can be helpful in our task because it highlights the decisive image region that is easily overlooked without highlighting. Left: aircraft. Right: cat.

is unlikely that currently predominant explanation approaches for visual tasks can solve this issue, as they are conceptually unsuitable to help decision makers detect AI mistakes. Overall, this suggests a paradigm shift away from the common "recommend-and-explain" approach may be necessary to effectively augment human decision-making in difficult decision tasks.

We see two promising broad directions for such a paradigm shift. One direction could be to reconsider the goal of explanations. Currently, most explanation approaches are technical in nature and aim to explain how the AI produced its outputs. Rather than enhancing people's decision-making process, these explanations divert attention toward trying to understand the AI model. Alternatively, explanations could provide information that naturally fits into people's decision-making process, which is the aim of *human-centered XAI* [86]. For example, inspired by how clinicians validate their colleagues' suggestions, Yang et al. [87] provided references to biomedical literature as explanations for AI recommendations. Ehsan et al. [88] explored the inclusion of socio-organizational context into AI explanations in a sales application. Such human-centered explanations may be more helpful in challenging decisions as they provide information that is relevant to the decision task itself. This may improve decision makers' ability to form an opinion of their own and to reconcile it with the AI's recommendation. This is in contrast to classical technical explanations, where in case decision makers have no clue about the correct answer, they have nothing more to work with than the AI's recommendation, which easily results in overreliance, as in our experiment.

Another promising approach is to allocate roles to AI other than providing decision recommendations, as recently proposed by some authors [89,90,91,92]. A good example of this approach for a visual decision task was provided by Lindvall et al. [20] in the context of cancer assessments. Their system helped pathologists to quickly identify and navigate to image regions of interest that they can review, without explicitly suggesting

whether the image contains cancer tissue or not. To an extent, such a design evades the issue of appropriate reliance, as it does not make users wonder whether they should rely on an AI recommendation or not. Instead, the system helps users to solve the task themselves.

6.4. Limitations

We caution against overgeneralizing our results, as we only explored a single task with a single explanation style and algorithm. A multitude of factors can have an influence on human reliance on AI, which is a fundamental challenge to the generalizability of empirical research in AI-assisted decision-making [93,94] that also applies to our work. Our study demonstrates that decision difficulty can be an important factor in human reliance behavior, but further studies are required to understand how our results apply to different settings. Moreover, to be able to use the appropriateness of reliance measurement concept by Schemmer et al. [14], we used a setup where participants made an initial decision independently from the AI. This workflow is known to decrease overreliance, at the cost of worse user experience [6,95]. On the other hand, overreliance in high mental workload decisions was high in our study despite this setup; the issue might be even more pronounced in a setup where humans directly receive AI support, which is often the case in real applications.

We used EEG to measure changes in mental workload in this study. While important for assessing the subjectively experienced challenge of a decision, this also imposed certain constraints on the decision task to avoid the introduction of artifacts into the EEG signal. For instance, we had to impose a fixed time interval for each decision instance, participants were required to move as little as possible, and we could not use text-based tasks. Future work could explore the usage of other physiological signals like electrocardiography, respiration, electrodermal activity, or blood pressure to measure mental workload [96]. Moreover, the use of EEG contributed to a gender imbalance in our results, as the application of conductive gel led to a limited number of female volunteers participating. Furthermore, EEG data has been obtained only once from each participant. While this is common in EEG-based studies, the reliability of the mental workload measure was only evaluated internally during the experimental period, not for other days or hours.

7. Conclusion

In our study, we monitored brain activity to understand how people's reliance on AI depends on how difficult a decision is for a person, as indicated by the mental workload induced by the decision. EEG signals were crucial to capture decision difficulty as subjectively experienced by participants, without resorting to potentially biased and distracting self-assessments from participants. We observed that participants tended to experience higher mental workload on task instances with higher pre-assigned difficulties, confirming that mental workload is a useful indicator of decision difficulty. We further found that the appropriateness of human reliance on AI depends strongly on the mental workload induced by the difficulty of a decision task, with people becoming much more overreliant in high mental workload decisions. While explanations had no significant effect in our study, the separate analysis of relative self- and AI reliance revealed the conceptual lim-

itation of saliency-based explanations to improve the appropriateness of reliance in high mental workload decisions. Our results imply that while the common "recommend-and-explain" approach to AI-assisted decision-making can effectively support humans on less challenging decision tasks, it may be less suitable for more ambitious applications like the interpretation of medical images. Future work should consider decision difficulty as an important factor in human reliance on AI and especially investigate how to effectively support human decision-making in difficult decisions.

Acknowledgments

The research reported in this work was partially supported by the EU H2020 ICT48 project "Humane AI Net" under contract # 952026. The support is gratefully acknowledged.

Disclosure of Interest

The authors report there are no competing interests to declare.

References

[1] Raaijmakers S. Artificial intelligence for law enforcement: challenges and opportunities. IEEE Security & Privacy. 2019 Sep;17(5):74-7. Available from: https://ieeexplore.ieee.org/abstract/document/8821442.

[2] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature Biomedical Engineering. 2018 Oct;2(10):719-31. Available from: https://www.nature.com/articles/s41551-018-0305-z.

[3] Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General. 2015;144(1):114-26. Available from: http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000033.

[4] Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. Journal of Marketing Research. 2019 Oct;56(5):809-25. Available from: https://doi.org/10.1177/0022243719851788.

[5] Prahl A, Van Swol L. Understanding algorithm aversion: when is advice from automation discounted? Journal of Forecasting. 2017;36(6):691-702. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2464.

[6] Buçinca Z, Malaya MB, Gajos KZ. 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. 2021 Apr;5(CSCW1):188:1-188:21. Available from: http://arxiv.org/abs/2102.09692.

[7] Green B, Chen Y. The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction. 2019 Nov;3(CSCW):50:1-50:24. Available from: https://doi.org/10.1145/3359152.

[8] Chiang CW, Yin M. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In: Proceedings of the 13th ACM Web Science Conference 2021. WebSci '21. Virtual Event, United Kingdom: ACM; 2021. p. 120-9. Available from: https://dl.acm.org/doi/10.1145/3447535.3462487.

[9] Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. Yokohama, Japan: ACM; 2021. p. 81:1-81:16. Available from: https://dl.acm.org/doi/10.1145/3411764.3445717.

[10] Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction. 2023 Apr;7(CSCW1):129:1-129:38. Available from: https://dl.acm.org/doi/10.1145/3579605.

[11] Yang F, Huang Z, Scholtz J, Arendt DL. How do visual explanations foster end users' appropriate trust in machine learning? In: Proceedings of the 25th International Conference on Intelligent User Interfaces. IUI '20. Cagliari, Italy: ACM; 2020. p. 189-201. Available from: https://dl.acm.org/doi/10.1145/3377325.3377480.

[12] Wang X, Yin M. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In: Proceedings of the 26th International Conference on Intelligent User Interfaces. IUI '21. College Station, TX, USA: ACM; 2021. p. 318-28. Available from: https://mingyin.org/paper/IUI-21/iui21.pdf.

[13] Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry. 2021 Jun;11(1):108:1-108:9. Available from: http://www.nature.com/articles/s41398-021-01224-x.

[14] Schemmer M, Kuehl N, Benz C, Bartos A, Satzger G. Appropriate reliance on AI advice: conceptualization and the effect of explanations. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. IUI '23. Sydney, NSW, Australia: ACM; 2023. p. 410-22. Available from: https://dl.acm.org/doi/10.1145/3581641.3584066.

[15] Bussone A, Stumpf S, O'Sullivan D. The role of explanations on trust and reliance in clinical decision support systems. In: Proceedings of the 2015 International Conference on Healthcare Informatics. ICHI 2015. Dallas, TX, USA: IEEE; 2015. p. 160-9. Available from: http://ieeexplore.ieee.org/document/7349687/.

[16] Lai V, Tan C. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19. Atlanta, GA, USA: ACM; 2019. p. 29-38. Available from: https://doi.org/10.1145/3287560.3287590.

[17] Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. Yokohama, Japan: ACM; 2021. p. 237:1-237:52. Available from: https://dl.acm.org/doi/10.1145/3411764.3445315.

[18] Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. Honolulu, HI, USA: ACM; 2020. p. 589:1-589:12. Available from: https://dl.acm.org/doi/10.1145/3313831.3376718.

[19] Van Berkel N, Ahmad OF, Stoyanov D, Lovat L, Blandford A. Designing visual markers for continuous artificial intelligence support: a colonoscopy case study. ACM Transactions on Computing for Healthcare. 2021 Jan;2(1):7:1-7:24. Available from: https://dl.acm.org/doi/10.1145/3422156.

[20] Lindvall M, Lundström C, Löwgren J. Rapid assisted visual search: supporting digital pathologists with imperfect AI. In: Proceedings of the 26th International Conference on Intelligent User Interfaces. IUI '21. College Station, TX, USA: ACM; 2021. p. 504-13. Available from: https://dl.acm.org/doi/10.1145/3397481.3450681.

[21] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021 Nov;3(11):e745-50. Available from: https://doi.org/10.1016/S2589-7500(21)00208-9.

[22] Lu Z, Yin M. Human reliance on machine learning models when performance feedback is limited: heuristics and risks. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. Yokohama, Japan: ACM; 2021. p. 78:1-78:16. Available from: https://dl.acm.org/doi/10.1145/3411764.3445562.

[23] Papenmeier A, Kern D, Hienert D, Kammerer Y, Seifert C. How accurate does it feel? – human perception of different types of classification mistakes. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22. New Orleans, LA, USA: ACM; 2022. p. 180:1-180:13. Available from: https://dl.acm.org/doi/10.1145/3491102.3501915.

[24] Parkes A. The effect of individual and task characteristics on decision aid reliance. Behaviour

& Information Technology. 2017 Feb;36(2):165-77. Available from: https://doi.org/10.1080/0144929X.2016.1209242.

[25] Longo L, Wickens CD, Hancock G, Hancock PA. Human mental workload: A survey and a novel inclusive definition. Frontiers in psychology. 2022;13:883321.

[26] Hancock PA, Meshkati N, Robertson M. Physiological reflections of mental workload. Aviation, space, and environmental medicine. 1985;56(11):1110-4.

[27] Backs RW. Going beyond heart rate: autonomic space and cardiovascular assessment of mental workload. The international journal of aviation psychology. 1995;5(1):25-48.

[28] Miller S. Workload measures. National Advanced Driving Simulator Iowa City, United States. 2001.

[29] Hirshfield LM, Chauncey K, Gulotta R, Girouard A, Solovey ET, Jacob RJ, et al. Combining electroencephalograph and functional near infrared spectroscopy to explore users' mental workload. In: Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: 5th International Conference, FAC 2009 Held as Part of HCI International 2009 San Diego, CA, USA, July 19-24, 2009 Proceedings 5. Springer; 2009. p. 239-47.

[30] Miller MW, Rietschel JC, McDonald CG, Hatfield BD. A novel approach to the physiological measurement of mental workload. International Journal of Psychophysiology. 2011;80(1):75-8.

[31] Hogervorst MA, Brouwer AM, Van Erp JB. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. Frontiers in neuroscience. 2014;8:322.

[32] Murata A. An attempt to evaluate mental workload using wavelet transform of EEG. Human Factors. 2005;47(3):498-508.

[33] So WK, Wong SW, Mak JN, Chan RH. An evaluation of mental workload with frontal EEG. PloS one. 2017;12(4):e0174949.

[34] Gevins A, Smith ME. Neurophysiological measures of cognitive workload during human-computer interaction. Theoretical issues in ergonomics science. 2003;4(1-2):113-31.

[35] Kramer AF. Physiological metrics of mental workload: A review of recent progress. Multiple task performance. 2020:279-328.

[36] Brookings JB, Wilson GF, Swain CR. Psychophysiological responses to changes in workload during simulated air traffic control. Biological psychology. 1996;42(3):361-77.

[37] Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neuroscience & Biobehavioral Reviews. 2014;44:58-75.

[38] Aricò P, Borghini G, Di Flumeri G, Colosimo A, Pozzi S, Babiloni F. A passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks. Progress in brain research. 2016;228:295-328.

[39] Borghini G, Aricò P, Di Flumeri G, Babiloni F, Borghini G, Aricò P, et al. Mental states in aviation. Industrial Neuroscience in Aviation: Evaluation of Mental States in Aviation Personnel. 2017:29-56.

[40] Gevins A, Smith ME, McEvoy L, Yu D. High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. Cerebral cortex (New York, NY: 1991). 1997;7(4):374-85.

[41] Missonnier P, Deiber MP, Gold G, Millet P, Gex-Fabry Pun M, Fazio-Costa L, et al. Frontal theta event-related synchronization: comparison of directed attention and working memory load effects. Journal of neural transmission. 2006;113:1477-86.

[42] Stipacek A, Grabner R, Neuper C, Fink A, Neubauer A. Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load. Neuroscience letters. 2003;353(3):193-6.

[43] Sweller J. Cognitive load during problem solving: Effects on learning. Cognitive science. 1988;12(2):257-85.

[44] Paas F, Renkl A, Sweller J. Cognitive load theory and instructional design: Recent developments. Educational psychologist. 2003;38(1):1-4.

[45] Yu K, Prasad I, Mir H, Thakor N, Al-Nashash H. Cognitive workload modulation through degraded visual stimuli: A single-trial EEG study. Journal of neural engineering. 2015;12(4):046020.

[46] Gupta SS, Manthalkar RR. Classification of visual cognitive workload using analytic wavelet transform. Biomedical Signal Processing and Control. 2020;61:101961.

[47] Kartali A, Janković MM, Gligorijević I, Mijović P, Mijović B, Leva MC. Real-time mental workload estimation using eeg. In: Human Mental Workload: Models and Applications: Third International Symposium, H-WORKLOAD 2019, Rome, Italy, November 14–15, 2019, Proceedings 3. Springer; 2019. p.

20-34.

[48] Steyvers M, Tejeda H, Kerrigan G, Smyth P. Bayesian modeling of human–AI complementarity. Proceedings of the National Academy of Sciences. 2022 Mar;119(11):1-7. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.2111547119.

[49] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision. 2015 Dec;115(3):211-52. Available from: https://doi.org/10.1007/s11263-015-0816-y.

[50] Hemmer P, Westphal M, Schemmer M, Vetter S, Vössing M, Satzger G. Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. IUI '23. Sydney, NSW, Australia: ACM; 2023. p. 453-63. Available from: https://dl.acm.org/doi/10.1145/3581641.3584052.

[51] Tejeda Lemus H, Kumar A, Steyvers M. An empirical investigation of reliance on AI-assistance in a noisy-image classification task. In: Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence. vol. 354 of Frontiers in Artificial Intelligence and Applications. Amsterdam, The Netherlands: IOS Press; 2022. p. 225-37. Available from: https://ebooks.iospress.nl/doi/10.3233/FAIA220201.

[52] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2017. Honolulu, HI, USA: IEEE Computer Society; 2017. p. 2261-9. Available from: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.243.

[53] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al.. PyTorch: An imperative style, high-performance deep learning library. arXiv; 2019. Available from: http://arxiv.org/abs/1912.01703.

[54] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision. 2020 Feb;128(2):336-59. Available from: https://doi.org/10.1007/s11263-019-01228-7.

[55] Gildenblat J, contributors. PyTorch library for CAM methods. GitHub; 2021. https://github.com/jacobgil/pytorch-grad-cam.

[56] Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378-82. Available from: https://psycnet.apa.org/doi/10.1037/h0031619.

[57] Xiao R, Ding L. Evaluation of EEG features in decoding individual finger movements from one hand. Computational and mathematical methods in medicine. 2013;2013(1):243257.

[58] McFarland DJ, McCane LM, David SV, Wolpaw JR. Spatial filter selection for EEG-based communication. Electroencephalography and clinical Neurophysiology. 1997;103(3):386-94.

[59] Delorme A. EEG is better left alone. Scientific reports. 2023;13(1):2372.

[60] Bajaj N, Carrión JR, Bellotti F, Berta R, De Gloria A. Automatic and tunable algorithm for EEG artifact removal using wavelet decomposition with applications in predictive modeling during auditory tasks. Biomedical Signal Processing and Control. 2020;55:101624.

[61] Shafiei SB, Shadpour S, Shafqat A. Mental workload evaluation using weighted phase lag index and coherence features extracted from EEG data. Brain research bulletin. 2024:110992.

[62] Luque F, Armada V, Piovano L, Jurado-Barba R, Santamaría A. Understanding pedestrian cognition workload in traffic environments using virtual reality and electroencephalography. Electronics. 2024;13(8):1453.

[63] Gu B, Chen L, Ke Y, Zhou Y, Yu H, Wang K, et al. The effects of varying levels of mental workload on motor imagery based brain-computer interface. International Journal of Embedded Systems. 2020;12(3):315-23.

[64] Roy RN, Charbonnier S, Campagne A, Bonnet S. Efficient mental workload estimation using task-independent EEG features. Journal of neural engineering. 2016;13(2):026019.

[65] Klimesch W. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain research reviews. 1999;29(2-3):169-95.

[66] Raufi B, Longo L. An evaluation of the EEG alpha-to-theta and theta-to-alpha band ratios as indexes of mental workload. Frontiers in Neuroinformatics. 2022;16:861967.

[67] Dan A, Reiner M, et al. Real time EEG based measurements of cognitive load indicates mental states during learning. Journal of Educational Data Mining. 2017;9(2):31-44.

[68] Mastropietro A, Pirovano I, Marciano A, Porcelli S, Rizzo G. Reliability of mental workload index assessed by eeg with different electrode configurations and signal pre-processing pipelines. Sensors.

2023;23(3):1367.

[69] Hjorth B. EEG analysis based on time domain properties. Electroencephalography and clinical neurophysiology. 1970;29(3):306-10.

[70] Hjorth B. An on-line transformation of EEG scalp potentials into orthogonal source derivations. Electroencephalography and clinical neurophysiology. 1975;39(5):526-30.

[71] Dao NA, Nguyen QA. Mental states detection by extreme gradient boosting and k-means. In: Deep Learning and Other Soft Computing Techniques: Biomedical and Related Applications. Springer; 2023. p. 23-33.

[72] Louis LEL, Moussaoui S, Van Langhenhove A, Ravoux S, Le Jan T, Roualdes V, et al. Cognitive tasks and combined statistical methods to evaluate, model, and predict mental workload. Frontiers in Psychology. 2023;14:1122793.

[73] Garg S, Shukla UP, Cenkeramaddi LR. Detection of depression using weighted spectral graph clustering with EEG biomarkers. IEEE Access. 2023;11:57880-94.

[74] Yamada Y, Kobayashi M. Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. Artificial intelligence in medicine. 2018;91:39-48.

[75] Zhang J, Cui X, Li J, Wang R. Imbalanced classification of mental workload using a cost-sensitive majority weighted minority oversampling strategy. Cognition, Technology & Work. 2017;19:633-53.

[76] Maddirala AK, Veluvolu KC. Eye-blink artifact removal from single channel EEG with k-means and SSA. Scientific Reports. 2021;11(1):11043.

[77] Wang G, Yin Z, Zhao M, Tian Y, Sun Z. Identification of human mental workload levels in a language comprehension task with imbalance neurophysiological data. Computer Methods and Programs in Biomedicine. 2022;224:107011.

[78] Al-Mohair HK, Saleh JM, Suandi SA. Hybrid human skin detection using neural network and k-means clustering technique. Applied Soft Computing. 2015;33:337-47.

[79] Shaheen N, Raza B, Shahid AR, Alquhayz H. A novel optimized case-based reasoning approach with K-means clustering and genetic algorithm for predicting multi-class workload characterization in autonomic database and data warehouse system. IEEE Access. 2020;8:105713-27.

[80] Hosseini MP, Hosseini A, Ahi K. A review on machine learning for EEG signal processing in bioengineering. IEEE reviews in biomedical engineering. 2020;14:204-18.

[81] Kamthekar S, Iyer B. A review of unsupervised learning algorithms for EEG signal analysis in the emotion detection applications. Available at SSRN 4291749. 2022.

[82] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019 May;1(5):206-15. Available from: `https://www.nature.com/articles/s42256-019-0048-x`.

[83] Fok R, Weld DS. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. AI Magazine. 2024 Jul;(Early View):1-16. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12182`.

[84] Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18. Montreal, Canada: ACM; 2018. p. 582:1-582:18. Available from: `http://dl.acm.org/citation.cfm?doid=3173574.3174156`.

[85] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018 Sep;6:52138-60. Available from: `https://ieeexplore.ieee.org/document/8466590/`.

[86] Ehsan U, Wintersberger P, Liao QV, Watkins EA, Manger C, Daumé III H, et al. Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems. CHI EA '22. New Orleans, LA, USA: ACM; 2022. p. 109:1-109:7. Available from: `https://doi.org/10.1145/3491101.3503727`.

[87] Yang Q, Hao Y, Quan K, Yang S, Zhao Y, Kuleshov V, et al. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Hamburg, Germany: ACM; 2023. p. 14:1-14:14. Available from: `https://dl.acm.org/doi/10.1145/3544548.3581393`.

[88] Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD. Expanding explainability: towards social transparency in AI systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. Yokohama, Japan: ACM; 2021. p. 82:1-82:19. Available from: `https://dl.acm.org/doi/10.1145/3411764.3445188`.

[89] Miller T. 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. Chicago, IL, USA: ACM; 2023. p. 333-42. Available from: https://dl.acm.org/doi/10.1145/3593013.3594001.

[90] Zhang ZT, Liu Y, Hussmann H. Forward reasoning decision support: toward a more complete view of the human-AI interaction design space. In: CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter. CHItaly '21. Bolzano, Italy: ACM; 2021. p. 18:1-18:5. Available from: https://doi.org/10.1145/3464385.3464696.

[91] Koon S. A human-capabilities orientation for human-AI interaction design. In: Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22). Virtual Event, USA; 2022. p. 1-5.

[92] Buçinca Z, Chouldechova A, Wortman Vaughan J, Gajos KZ. Beyond end predictions: stop putting machine learning first and design human-centered AI for decision support. In: Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22). Virtual Event, USA; 2022. p. 1-4.

[93] Lai V, Chen C, Smith-Renner A, Liao QV, Tan C. 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. Chicago, IL, USA: ACM; 2023. p. 1369-85. Available from: https://dl.acm.org/doi/10.1145/3593013.3594087.

[94] Salimzadeh S, Gadiraju U. "DecisionTime": a configurable framework for reproducible human-AI decision-making studies. In: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. UMAP Adjunct '24. Cagliari, Italy: ACM; 2024. p. 66-9. Available from: https://doi.org/10.1145/3631700.3664885.

[95] Fogliato R, Chappidi S, Lungren M, Fisher P, Wilson D, Fitzke M, et al. 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. Seoul, Republic of Korea: Association for Computing Machinery; 2022. p. 1362-74. Available from: https://dl.acm.org/doi/10.1145/3531146.3533193.

[96] Charles RL, Nixon J. Measuring mental workload using physiological measures: A systematic review. Applied ergonomics. 2019;74:221-32.

[97] Leifeld P. texreg: Conversion of statistical model output in R to LaTeX and HTML tables. Journal of Statistical Software. 2013;55(8):1-24. Available from: https://doi.org/10.18637/jss.v055.i08.

A. Logistic Regression Models

Table 2. Logistic regression models used in Section 5.2. Table produced with texreg [97].

	RSR	RAIR	Acc low MWL	Acc high MWL
(Intercept)	2.19^{***}	0.18	2.45^{***}	0.89^{***}
	(0.30)	(0.17)	(0.13)	(0.08)
condition_contrast	0.03	−0.01	−0.06	0.13
	(0.25)	(0.16)	(0.14)	(0.09)
MWL_contrast	-2.29^{***}	−0.24		
	(0.49)	(0.18)		
condition_contrast:MWL_contrast	−0.71	−0.10		
	(0.51)	(0.33)		
AIC	571.07	1468.69	1481.20	3246.09
BIC	602.92	1503.78	1498.59	3263.80
Log Likelihood	−278.54	−727.34	−737.60	−1620.04
Num. obs.	699	1111	2437	2706
Num. groups: Participant_ID_Numeric	33	33	33	33
Var: Participant_ID_Numeric (Intercept)	1.20	0.67	0.35	0.16
Var: Participant_ID_Numeric MWL_contrast	1.04	0.05		
Cov: Participant_ID_Numeric (Intercept) MWL_contrast	−0.93	−0.18		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$