Confidence-based deferral time in joint human-AI decision-making

Testing a novel approach

Schmidt, Mykel J.A.

Award date:
2023

Eindhoven, 07/05/2023

Confidence-based deferral time in joint human-AI decision-making

Testing a novel approach

by Mykel John Alexander Schmidt

identity number: 0912900

in partial fulfilment of the requirements for the degree of

Master of Science
in Human-Technology Interaction

Supervisors:
dr. Chao Zhang
dr. ir. R. Conijn

Abstract

The usage of AI throughout society is rapidly increasing and it becomes more and more important to understand how to optimize joint human-AI decision-making processes. To this end, this study investigates two ways of incorporating AI decision confidence in joint human-AI decision-making processes, one being a completely novel approach. Both approaches relate to the deferral method; a method of human-AI decision-making in which the AI takes an initial decision and subsequently defers the decision to the human who makes the final decision. The novel approach involves using the confidence level of the AI's decision to determine the deferral time. The system will give the human decision-maker more time to decide when the AI is uncertain and less time to decide when the AI is certain. The other approach is presenting the confidence of the AI together with its decision. The two approaches are tested in an online experiment with 266 subjects who were randomly distributed across four between-subject conditions. The results indicate that presenting confidence information slows down the decision-making speed while not improving accuracy. Furthermore, the study found no evidence to support the effectiveness of a confidence-based deferral time in improving the accuracy or speed of joint human-AI decision-making. This study provides a valuable insight into the integration of AI prediction confidence in human-AI decision-making and highlights the need for further research in this area.

Keywords: joint human-AI decision-making, artificial intelligence, AI confidence, deferral time, confidence-based deferral time, XAI

Preface

I am pleased to present my master thesis which explores a novel concept in the realm of joint human-AI decision-making with the confidence-based deferral time. I would like to sincerely thank my supervisors dr. C. Zhang and dr. ir. R. Conijn for their guidance and support during the project. Their expertise, insights, and encouragement, have been very helpful in helping me guide through this project. I hope this thesis will make a meaningful contribution to this interesting field and I am grateful for the opportunity to contribute to it.

Contents

1. Introduction

1.1. Background of Human-AI Decision-making

Joint human-AI decision-making is a collaborative process in which the human together with an AI forms a decision. Human decision-making and AI decision-making both have their own strengths and weaknesses. For instance, humans are unable to process large amounts of data and parameters (Tegmark, 2018), are subjected to a long list of biases that are a field of study on their own (Kahneman et al., 1982), and are relatively slow at making decisions (Siegel & Sapru, 2005). Furthermore, humans' decisions are inconsistent; a huge number of factors can influence it, such as being tired, upset, or hungry (De Ridder et al., 2014). On the other hand, AI decision-making is typically more consistent, even more complex models such as deep neural networks that are stochastic in nature. This consistency can be a great strength. Furthermore, AI is able to make these decisions while processing huge amounts of data at lightning speed (Korteling et al., 2021), thereby tremendously reducing the decision time. However, AI decision-making isn't flawless either. AI lacks the ethical, moral, empathic, and emotional understanding that humans naturally have due to our shared human experience. Its lack of genuine empathy in their decisions, may lead to (unintended) harmful consequences for human-beings affected by the decision. Moreover, AI lacks human's "common sense", which underlie a lot of decisions humans make in day-to-day life.

In conclusion, from the above it is evident that the weaknesses of one are actually part of the strengths of the other, and vice versa; while humans lack in speed and are unable to process large amounts of data, they are capable of understanding the moral and empathic context that may be missed by the AI. It is therefore no surprise that human-AI collaboration is increasingly used throughout society in decision-making processes and is repeatedly demonstrated to

outperform both human decision-making and AI decision-making in various contexts, such as in medical decision-making (Reverberi et al., 2022).

1.2. Problem Statement

As joint human-AI decision-making is becoming increasingly prevalent in various domains, the optimization of the collaborative decision-making process is crucial for effective and efficient performance. To provide a better understanding of the problem statement, this section discusses two practical domains of joint human-AI decision-making in further depth.

Security screening is one such domain where joint human-AI decision-making is increasingly being applied. Currently, various airports are already implementing facial scanners. Heathrow airport is among these airports (Kobie, 2018). When a passenger shows up at check-in, the system will take a digital image of their face, compare it to the one on their scanned passport, and tie it to their flight details. However, in situations where the AI system is unsure, a manual check by a supervisor is necessary. This is an example of joint human-AI decision-making, where the AI takes the brunt of the decisions and defers the decision to manual operators when it is uncertain. Due to the huge volume of incoming and outcoming passengers, it is a great area for optimization, as an increase in joint human-AI decision-making speed will yield great benefits for the airport, especially since a short-staffing problem has been reported in this domain (Kobie, 2018).

Another prominent domain where effective and efficient performance of joint human-AI decision-making is of much value, is the domain of content moderation. The use of social media is growing and growing. As a result, the need for digital content moderation on social platforms is ever-increasing. Some statistics are as follows; everyday, 95 million photos and videos are uploaded to Instagram (Cveticanin, 2023), over 4000 photos are uploaded per second to

Facebook (TrueList, 2023), and close to a billion tweets are made every day (Petrov, 2023). This huge incoming stream of content needs to be moderated 24/7 for a platform to conform to their community guidelines. This includes analyzing a wide range of content, from text to visual content. Even within the content type, there are a wide range of potential rule violations. For instance, for textual content, it could include bullying, propaganda, misinformation, threats, and so on. Fortunately, AI can play a prominent role in content moderation through the use of various forms of AI: natural language processing, image recognition, computer vision, and computer audition. Unfortunately, the AI is not faultless and thus manual reviewing of the content remains necessary. Therefore, employees are hired to do manual moderation of the content that has been selected by the AI for potential rule violations. This is another example of joint human-AI decision-making in which the AI defers decisions to the human. Unfortunately, for the employee, manual content moderation often comes with a damaging psychological impact and a heavy workload due to the large amount of incoming content. It is therefore an important area for optimization with respect to human-AI decision-making. An increase in joint human-AI decision-making speed will shorten the total time that needs to be spent on content moderation, which is beneficial for a company that has to moderate huge streams of incoming content. Furthermore, as with any other domain, an increase in decision accuracy is always considered beneficial. In this instance, the increase in joint human-AI decision-making accuracy will improve the platform's consistency with respect to their community guidelines.

The two practical domains share a similarity in that both domains perpetually deal with a huge volume of incoming decisions. Therefore, besides decision accuracy, the decision speed (or efficiency) of the collaborative decision-making process is another parameter in which

improvements are valuable. In conclusion, these two practical domains highlight the importance of the parameters decision speed and decision accuracy in joint human-AI decision-making.

1.3. Research Objective

There are various ways that improvements on joint human-AI decision-making speed or accuracy can be realized. This study focuses on two methods that have the potential to do so. Both methods relate to the manner in which the confidence of the AI's initial decision is communicated to the human decision-maker (to whom the decision is deferred). The first method, dubbed 'confidence presentation', simply consists of displaying the AI's confidence as a percentage value between 0 and 100 to the human decision-maker. The second method, dubbed 'confidence-based deferral time', works by adjusting the time interval the human decision-maker has to make the decision. In this proposed deferral framework, the human gets more time when the AI is unsure of its initial decision, but less time when the AI is more certain of its initial decision.

In conclusion, the objective of this research is to explore the effects on the accuracy and speed of joint human-AI decision-making of two particular methods: confidence presentation and the usage of a confidence-based deferral time.

1.4. Relevance of the Research

As highlighted in the problem statement, in various domains where joint human-AI decision-making is applied—particularly in those where there is a perpetual and huge influx of decisions to be made—there is a high value in improving the decision speed and accuracy. Namely, the improvement in these parameters of joint human-AI decision-making leads to various beneficial outcomes depending on the domain and context (e.g., increased security and decreased costs).

From a scientific perspective, this research is primarily related to the field of Human-Technology Interaction and furthermore to the fields of Artificial Intelligence and Cognitive Psychology (Decision-Making). With respect to the scientific relevance, this research will further deepen the scientific knowledge regarding how the AI should communicate the confidence of its decisions to the human, and how, resultantly, they together can make optimal decisions. By exploring the effects of confidence presentation and the novel approach of the confidence-based deferral time, this study will contribute new knowledge to the existing body of scientific literature surrounding this topic, and consequently help inform the design of future human-AI decision-making systems.

1.5. Thesis Outline

The thesis is structured as follows: First the literature will be reviewed which identifies current gaps in the literature that need to be addressed. This logically flows into the formulation of the research question and the hypotheses. Following that, the setup of the study will be discovered in great depth so potential replication will be possible. After that, the results of the study are shared, which is followed by an in depth discussion that includes the interpretation and the contextualization with earlier findings. Finally, suggestions for future research are provided and the thesis resolves with a final conclusion summarizing the main findings of this study.

2. Literature Review

This section includes a review of the current scientific literature regarding the topic of human-AI decision-making, explainable AI (in particular confidence presentation), and finally, time pressure and its effect on human decision-making. The section closes off with a conclusion of the literature review, which summarizes the findings and identifies the gaps, which logically leads to the next section; the research questions and hypotheses.

2.1. Joint Human-AI Decision-making

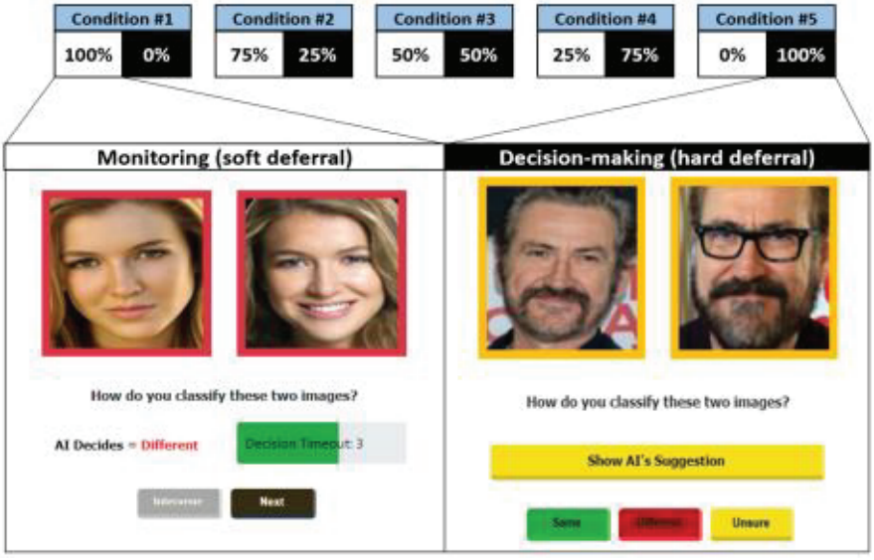
There are many forms of human-AI collaboration, and they can be categorized in a myriad of ways. For example, Dellermann et al. (2021) developed a taxonomy identifying 50 categories by evaluating each hybrid intelligence system based on task characteristic, learning paradigm, AI-human interaction, and human-AI interaction. In the context of decision-making, a simple way of differentiating these hybrid systems is by whether the human makes the decision never, sometimes, or always. When the AI makes each decision, the human's involvement in the decision-making process stems from initializing the AI's configuration and subsequent fine-tuning (e.g., autonomous systems). When the human makes each decision, the AI's involvement in the decision-making process stems from providing information or feedback on the decisions (e.g., recommender systems, evaluative systems). When the decision is sometimes made by the AI and other times by the human, the interaction strategy is referred to as mixed-initiative (Allen et al., 1999). The research in this paper pertains to such systems, more specifically, to systems using the deferral framework. In this framework, an AI agent defers the decision to the human-decision-maker based on a probabilistic or distance-based threshold in classification models (Salehi et al., 2021). This approach reaps the benefits from automation, while preventing the potential consequences (Madras et al., 2018). Therefore, this approach is

ideal for situations where there is a large volume of decisions that can be taken in an automated way, such as content moderation and security screening.

Salehi et al. (2021) conducted an experiment using a face matching task to investigate the effects of different deferral frameworks. In their study, they distinguished between two deferral frameworks, which they named “soft deferral” and “hard deferral”. In “soft deferral”, the human’s role is to monitor the decisions of the AI and intervene when deemed necessary. In “hard deferral”, the human’s role is to make the decisions and request the AI’s suggestions when deemed necessary. They allocated participants to one of five conditions, which included the two deferral frameworks in varying degrees as seen in Figure 1. The results showed that higher “hard deferral” rates led to an increase in sensitivity (which reflects the accuracy of a participant’s decisions based on Signal Detection Theory (Lynn & Barret, 2014)), but to a decrease in efficiency and trust in the AI.

Figure 1

Image of the conditions of the experiment by Salehi et al. (2021)



2.2. Explainable AI

Explainable AI (XAI) refers to the development of artificial intelligence systems that are capable of providing understandable insights to the user on their workings. This is typically done by providing insights or offering transparency in how they reached certain decisions or outcomes. These systems can positively affect the collaboration with human-beings as they become more understandable and trustworthy to the user (Burkart & Huber, 2021). XAI is becoming increasingly important as the application of AI systems continues to spread throughout society, including in high-stakes domains, such as healthcare (Yala et al., 2019). Furthermore, with the continuing spread of AI throughout society, AI will increasingly come into contact with laymen who do not have a deep understanding of such complex systems, which further reinforces the need for XAI. The increase in trust or cooperation resulting from XAI can positively affect the accuracy or the speed of joint human-AI decision-making by, for instance, decreasing second-guessing. Furthermore, the increase in the human's understanding of the AI system can improve accuracy of joint human-AI decision-making as well, as it can become clearer to the human decision-maker when the model is providing accurate predictions and when it is not.

Research by Lai et al. (2021) categorizes AI assistance elements in four broad categories; [1] predictions, [2] local explanations: information about predictions (which informs the user whether they should follow a particular prediction), [3] global explanations: information about the model (which provides the user an overall understanding of the model), and lastly, [4] other AI system elements affecting the user experience and agency. Within these categories, there are many different methods of assistance. For example, information about a prediction could include providing the rules the AI has followed to come to the prediction or a confidence metric attached

to the prediction. While information about the model could include information about the training data, or an overall metric of model performance (e.g., accuracy).

Various studies have been conducted on the effects of XAI elements with respect to human-AI decision-making. Research by Buçinca et al. (2021) finds that people often over rely on AI systems and accept their suggestions even when they are incorrect. However, they find that adding explanations or an indication of the AI's confidence does not necessarily reduce this tendency. Additionally, they also test three conditions with cognitive forcing interventions to force people to engage more thoughtfully with the AI-generated explanations. Although they find that these interventions do in fact reduce overreliance, people rated these conditions the least favorable subjectively. Furthermore, findings by research of Bansal et al. (2021), reaffirms that explanations increase the chance that humans will accept the AI's recommendation, regardless of its correctness, and hence do nothing to reduce the overreliance. This is echoed by research of Bussone et al. (2015), who researched XAI in the context of healthcare. Their research found that when the system gave a fuller explanation of the facts used in making a diagnosis, it had a positive effect on trust but also led to overreliance issues.

Hence, there is a need for XAI approaches that encourage appropriate trust calibration in AI, and not simply enhance trust. In the context of joint human-AI decision-making, trust calibration refers to appropriately adjusting one's level of trust in the AI system based on its capabilities and limitations (Zhang et al., 2020). The distinction between trust and trust calibration is critical. While enhancing trust is of importance as it allows people to be influenced by the AI's judgment, there are cases in which the user should actually not trust the AI's prediction and should remain more critical and rely more on their own judgment. Hence, for the sake of joint human-AI decision-making accuracy, trust calibration is a crucial aspect.

2.2.1. Confidence Presentation

One form of XAI that may combat overreliance, is the inclusion of confidence with the predictions of the AI. In this manner, the AI can communicate how reliable its predictions are to the user. This form of XAI is considered a local explanation as it provides information about a specific prediction (Lai et al., 2021). In theory, by including confidence with the AI's predictions, the human will know better when to follow the AI's decision, and when to be more critical. As a result, the overreliance would be reduced, in theory.

Research by Zhang et al. (2020) corroborates this. By conducting two human experiments, Zhang et al. find that confidence presentation can indeed help calibrate people's trust in an AI model. However, they find that trust calibration alone is not enough to improve the accuracy of joint human-AI decision-making, which also depends on whether the human and the AI can complement each other's performance. Furthermore, research by Yin et al. (2019), finds that people's trust in an AI is affected by both the stated accuracy (i.e., the overall accuracy of the AI model presented to the participant prior to the task) as well as the observed accuracy (i.e., the overall accuracy of the AI model in practice as observed by the participant). Furthermore, they find that the effect of the stated accuracy can change depending on the observed accuracy.

However, some research suggests that confidence presentation does not aid trust calibration. For instance, research by Buçinca et al. (2021) finds that stating the accuracy does not help prevent overreliance of the user. Furthermore, research by Lai et al. (2019) finds that participants are not sensitive to statements of machine accuracy and are more likely to trust machine predictions with an accuracy statement than without, even when that accuracy statement suggests that it is not being accurate.

Furthermore, not all AI models provide well-calibrated confidence scores. If a neural network predicts something with a probability of 0.3, this prediction should have a 30% chance of being correct, however, this is not always the case as many neural networks are prone to overconfidence (Wei et al., 2022). In addition, confidence can be communicated in a variety of ways (visual, percentual, decimal) under a variety of semantically related terms (certainty, reliableness, confidence, probability, likelihood). This freedom leads to ambiguity which can threaten the consistency between the interpretations of human decision-makers.

Finally, research by Miller (2019) regarding XAI, states that referring to probabilities or statistical relationships in explanations is not an effective approach. According to Miller, human decision-makers desire causal explanations for the AI's predictions.

In conclusion, confidence presentation may help increase the user's trust in the system. However, this alone does not necessarily lead to improved accuracy, particularly when the AI has a high error rate. In such cases, proper trust calibration is essential. However, findings are mixed with respect to the effectiveness of confidence presentation in properly communicating the reliability of the model's prediction, i.e., trust calibration. Therefore, the effect of confidence presentation on accuracy is still unresolved. In addition, there is a lack of a de facto standard of communicating AI confidence to the human. All in all, this suggests a need for more research surrounding the communication of confidence from the AI to the human.

2.3. Confidence-based Deferral Time

One novel approach of communicating the confidence of the AI is through the usage of a confidence-based deferral time. In this framework, the system provides the human decision-maker with more time to make the decision when the decision of the AI has more uncertainty. Conversely, when the AI's decision has more certainty, the system will provide the

human decision-maker with less time. In theory, this could help limit the amount of time and cognitive resources that are spent on decisions where the AI is already quite certain, and instead allocate that time and those cognitive resources to decisions where the AI is more uncertain. Research in cognitive psychology has established that the human brain has limited cognitive resources, which are required for processing information and making decisions, especially in complex and uncertain situations (Coxon, 2012). Theoretically, this novel approach could improve the accuracy of decisions the AI considers difficult, as the human decision-maker is able to dedicate more time and cognitive resources to those decisions. Hence, in theory, this deferral approach could result in more efficient and more accurate collaborative decision-making. Currently, there is no literature on this exact approach. This deferral approach is a new way of communicating confidence but may come with certain downsides, such as an increase in time pressure. Hence, the following section will review literature related to that phenomenon.

2.4. Time Pressure

The relevance of time pressure lies in its potential impact on the accuracy of joint human-AI decision-making. Research by Dambacher and Hübner (2015) looks at how time pressure affects decision-making between conflicting decisions. Their research found that response boundaries lower as time pressure increases. Response boundaries refer to the threshold of evidence or certainty that needs to be reached to trigger a decision. When time pressure increases, people tend to lower this threshold, and as a result, the accuracy may decrease. This is also known as the speed-accuracy tradeoff and the tradeoff has been studied and found across a wide variety of studies (Donkin et al., 2014). However, in addition to the strategic lowering of the response boundaries, which is a well-known reaction to time pressure, Dambacher and

Hübner also found that time pressure impaired early sensory filtering, which lowers the processing efficiency in decision-making.

Research by Rae et al. (2014), similarly finds that emphasizing decision speed over accuracy leads to a lowering of the response boundaries. In addition, they find that this shift of emphasis also leads to a decrease of the quality of the information being accumulated during the decision process.

In conclusion, these findings suggest that an increase in time pressure may not only lead to the strategic lowering of the evidence threshold by the decision-maker; it also affects the processing quality of the evidence by the decision-maker. Both of these factors are negatively impacting the accuracy of the decision-making process.

2.5. Conclusion

From the literature review various findings emerge. Namely, the presentation of confidence can enhance and calibrate trust, which may lead to better joint human-AI decision-making performance. However, other findings demonstrated it does not prevent overreliance and may even increase it. Therefore, there remains a need for more research into the effectiveness of including the confidence with the AI's decisions. Furthermore, the concept of a confidence-based deferral time in joint human-AI decision-making has not been researched yet. Lastly, the literature showed that time pressure has a negative effect on the accuracy of decisions, which could play a role in the implementation of a confidence-based deferral time. To address these gaps, the next section will formulate the research questions that this research centers around.

3. Research questions and hypotheses

From the literature review, two primary gaps emerged. First of all, the findings on the effectiveness of confidence presentation have not been conclusive. Secondly, the benefits of the usage of a confidence-based deferral time have not been researched before. Hence, the following research questions have been formulated for this study:

RQ1: *, how does confidence presentation affect the accuracy and the speed #?

RQ2: **, how does a confidence-based deferral time affect the accuracy and the speed #?

To test the research questions the following set of hypotheses have been formulated.

H1a: *, confidence presentation will increase the accuracy #

H1b: *, confidence presentation will decrease the speed #

H2a: **, a confidence-based deferral time will increase the accuracy #

H2b: **, a confidence-based deferral time will increase the speed #

* = Compared to no confidence presentation,

** = Compared to a fixed deferral time,

= of joint human-AI decision-making

The rationale behind the direction of H1a is based on the finding by Zhang et al. (2020) that confidence information leads to an increased trust calibration, which consequently should lead to an increase in accuracy, provided the error boundaries of the human and the AI differ (i.e., they complement each others skills). The hypothesized decrease in speed is due to an increase in information/evidence that needs to be processed by the human decision-maker.

As evident from the literature section of the confidence-based interval, the directions in H2 are self-hypothesized since there has been no previous research conducted on this particular deferral approach. The hypothesized increase in accuracy and speed stems from the reallocation

of cognitive resources (Coxon, 2012) and time by the human decision-maker to better fit the mental and temporal demands of each separate decision, i.e., allocating more resources for difficult decisions (increasing accuracy) and allocating less time for easier decisions (increasing speed).

In addition, the study also includes three subjective constructs. Namely, perceived workload, perceived time pressure, and system satisfaction. These constructs have been added to provide more understanding and context to the primary research findings.

4. Method

4.1. Study Design

The experiment has four between-subject conditions to which participants are randomly allocated. The first condition primarily serves as a baseline to determine the difficulty of the task for humans. The decision task is a binary decision task where for each of the 40 trials only one of two decision options is correct, this is explained in furth depth in 4.4 ‘Task Design’.

Conditions:

1. No AI assistance, Fixed (deferral) time
2. AI assistance, Fixed deferral time
3. AI assistance, Fixed deferral time, Confidence presentation
4. AI assistance, Confidence-based deferral time

AI assistance means that during the decision task the participant will be presented with the decision of the AI. This reflects the manner in which deferral systems work; the participant is deferred to by the AI with an initial decision of the AI. The AI has an average accuracy of 75% in the decision task. In condition 3, the AI’s decisions will be accompanied by a confidence percentage between 55% and 95% (a confidence of 50% would imply that the AI is guessing between the two options). In condition 4, the confidence-based deferral time is implemented. This means that the participant receives more or less time for the trial based on how confident the AI is in its decision for that trial. The participants are made aware of this mechanism. Therefore, they can interpret the time they receive as a communication of the AI’s confidence in its decision.

4.2. Participants

Participants are recruited using the Prolific.co platform (Prolific, n.d.). There are various prescreening requirements: participants are from the UK or the US, are fluent in English, and have a 97% or higher approval rate. The approval rate represents how often their previous experimental participations have been rejected or approved by other experiment leaders. Participants are paid £1.9 for their participation. In total, 276 participants' data were analyzed. The participants are between 18 and 45 years old with a mean age 31.9 years old. The recruitment filter of Prolific resulted in an equal distribution of male and female participants, with both genders comprising 50% of the total participants.

4.3. Procedure

Participants were directed from Prolific to the website where the experiment is hosted. There, they were greeted by a welcome screen with a description of the experiment. Once they clicked continue, they were shown a screen where they could give informed consent and permission for data collection. Furthermore, on this screen they were asked to fill in their Prolific ID. Following this screen, they were presented a new screen where they were given a detailed explanation of the decision task. This screen included a timer, so that participants could not simply continue to the next page before the timer ran out. This was to ensure participants read the description. Once participants had finished reading (and the timer had run out as well), they were able to press continue. Then they were presented with one more page before the decision task, this page included a short video of roughly 30 seconds that explained the decision task with more focus on the UI elements. Also here the page was given a timer so participants would watch the video. After this was done, participants could start the decision task. The decision task presented two photos that each depicted a face. The goal was to decide whether the two photos

belonged to the same individual or not. Depending on the condition, the participant also received input from the AI. Furthermore, the participant had a limited time to make the decision, which was made clear to them by a timer in the middle of the screen. The duration of the timer was dependent on the condition. After clicking “Same” or “Different”, the trial was completed and the participant was presented with a “Continue” button to continue to the next trial. The participant repeated this for a total of 40 trials. After completing 40 trials, the participant was presented with three questionnaires, measuring three subjective constructs, namely, perceived workload, time pressure, and finally, system satisfaction. Upon answering each item, the button “Finish experiment” was enabled and participants clicked it to be redirected to the final page that included the completion code needed for registering the completion of the experiment in Prolific. This completion page furthermore included a short debriefing that included their final accuracy for the decision task.

4.4. Task Design

The task consists of 40 trials. A trial consists of the presentation of two photos of which each photo depicts a face of an individual. As shown in Figure 2, the participant has a limited time to decide between two options. One option being that the photos depict the same individual, and the other option being that the photos depict different individuals. When the participant has clicked “Same” or “Different”, a trial is completed. The participant then has to click ‘Continue” to move on to the next trial.

The time the participant has to make the decision depends on the condition. In the fixed deferral time conditions (conditions 1 to 3), the participant has 13 seconds to make the decision for each trial. In the confidence-based deferral time condition (condition 4), the participant has between 9 and 17 seconds to make the decision (i.e., 13±4), depending on the confidence of the

AI for its decision during that trial. However, when averaging the decision time across all trials, condition 4 also has 13 seconds per trial. In other words, the sum of the decision time of all trials is the same across all conditions.

The decision to consider 13 seconds as the fixed deferral time was based on a combination of (informal) pre-testing and intuition. Naturally, the appropriate duration highly depends on the type of decision task and its context (e.g., complexity). After testing the options of 10 seconds and 15 seconds for the face-matching task, the former appeared to provide the decision-maker with insufficient time (only 6 seconds when the AI is the most confident), while the latter provided so much time even when the AI is the most confident (namely, 11 seconds) that the confidence-based manipulation may not really make a difference.

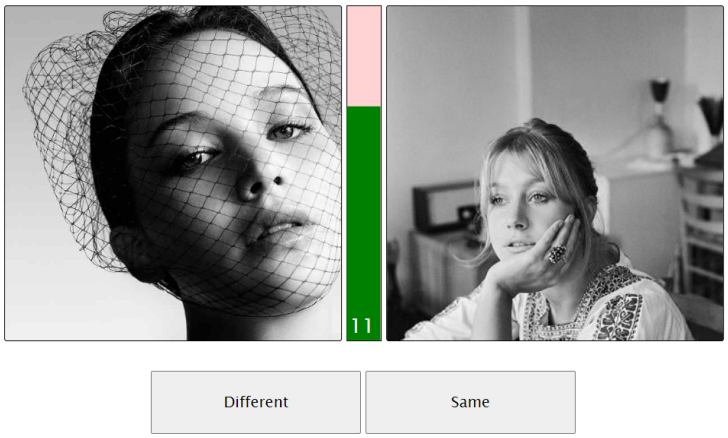
4.4.1. Implementation

The experiment is implemented as a web application, developed using HTML, CSS, and Javascript. The decision task only forms the middle part of the experiment. Prior to the decision task, participants are asked for informed consent and will receive information regarding the experiment. After the decision task, the participants fill in a questionnaire. The entire experiment runs on a single webpage. Javascript is used to hide and display elements to progress through the stages of the experiment and to record all the experimental data.

4.4.2. Manipulation and Interface Design

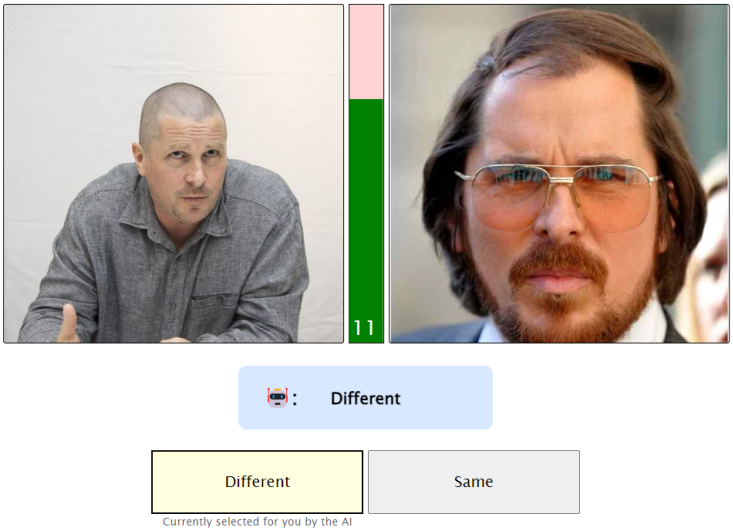
Condition 1. In condition 1, the participant makes all the decisions on their own without the help of an AI. Furthermore, the participant has a fixed time of 13 seconds. When they time out, the UI responds: “Failed to make a decision in time”.

Figure 2
Example of a participant in condition 1 (no-AI condition)



Condition 2. In condition 2, the participant will be deferred to by the AI with an initial decision. Furthermore, the participant has a fixed time of 13 seconds.

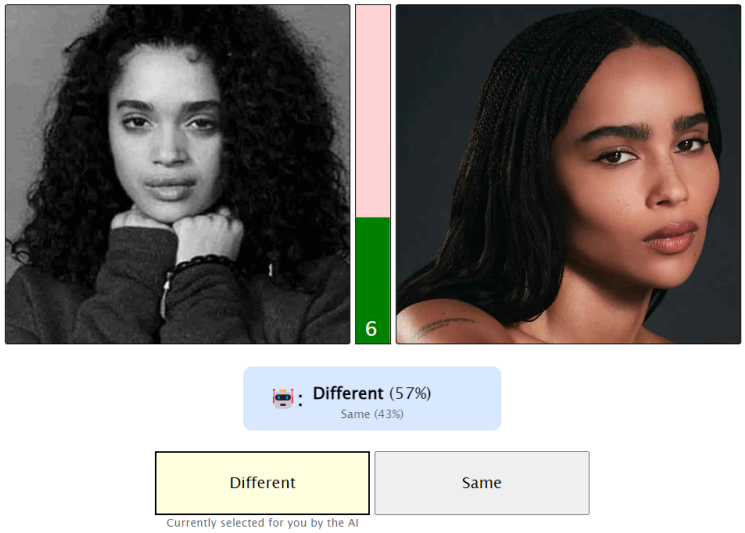
Figure 3
Example of a participant in condition 2 (basic AI condition)



Condition 3. In condition 3, the participant will also be deferred to by the AI with an initial decision. However, in this condition, the AI will also provide its confidence in its decision. Furthermore, like in the previous conditions, the participant has a fixed time of 13 seconds.

Figure 4

Example of a participant in condition 3 (confidence presentation condition)

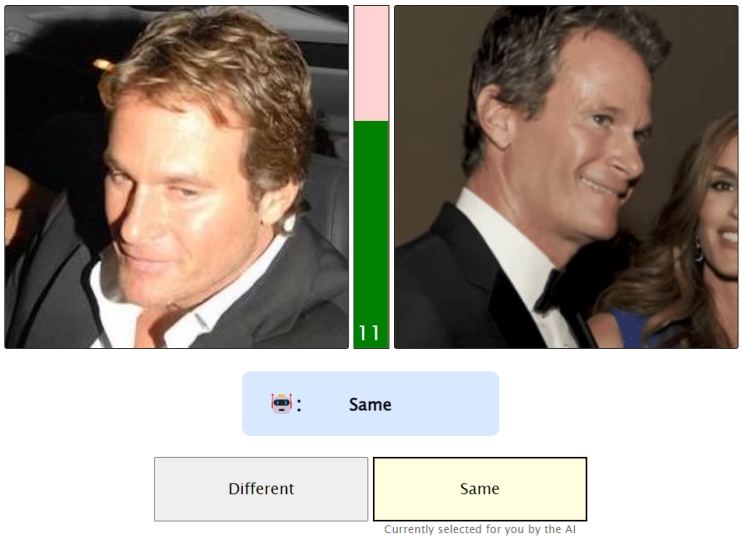


Condition 4. In condition 4, the participant will also be deferred to by the AI with an initial decision. However, in this condition, the system will provide more or less time based on how confident the AI is in its decision. The time of the 40 trials range between 9 and 17 (13±4). On average across all trials, all conditions offer the participants an equal amount of time. When the AI has a confidence of 95%, the system only gives 9 seconds. When the AI has a confidence of 55%, the system gives 17 seconds. All 40 trials fall within this confidence and time range. The relation between confidence and the time the system gives is linear and is captured with the following formula:

$$Time\ (seconds)\ =\ 8\ +\ ((100\ -\ confidence\%)\ \div\ 5)$$

Figure 5

Example of a participant in condition 4 (confidence-based deferral time condition)



Timing out. In condition 1, timing out will be met with a message: "Failed to make a decision in time". The missed trial does not affect their accuracy, as it is ignored (however, this is not made explicitly clear to the participants). In the AI conditions (2, 3, & 4), each trial starts off by the AI selecting a decision within the first second. When the participant does not change or accept the decision (by clicking "Same" or "Different") before the timer runs out, the system will take the initial decision of the AI as the final answer. That's why that button is highlighted yellow in those conditions.

4.4.3. Stimuli Selection

The decision task has a total of 40 trials. Therefore, 80 images are used in the experiment. Most of the stimuli are selected using VGGFace2; a dataset comprising millions of images of celebrities. The rest of the stimuli are downloaded manually by using Google Image Search. Each stimuli pair is verified by reverse image searching the photos (using Google Image Search and Yandex Image Search) and establishing the identities of the persons depicted.

4.4.4. Wizard-of-Oz AI

During the decision task, for each of the trials, the AI makes a decision with a certain confidence. However, the AI aspect is merely a simulation as this study employs a Wizard-of-Oz prototype as the AI. This is done to be able to control the performance of the AI. As will be further explained below, the accuracy of the AI for each participant is 75% and its mistakes and correct decisions are balanced in terms of decision classification (i.e., as many false positives as false negatives and as many true negatives as true positives).

To simulate the use of real AI, each stimuli pair is analyzed by an actual AI facial similarity model (ToolPie, n.d.). This model produces for each stimuli pair a similarity percentage. However, the experiment requires decisions of the AI in the experiment and the respective confidence for the decisions. Hence, the similarity scores of the AI model are mapped to the parameters needed for the experiment according to the matrix depicted in Table 1. This mapping process of actual AI similarity scores to the (Wizard-of-Oz) AI decisions and confidence percentages presented in the experiment was done in a methodical manner.

As the matrix in Table 1 shows, after the mapping, each of the stimuli pairs has an AI decision ("Same" or "Different") and a unique confidence percentage from a set of 40 values {55.5%, 56.6%, …, 94.5%}. As a result, in the experiment, the decision of the AI, as well as the confidence percentage the AI displays with its decision, has real meaning. In the set of stimuli pairs that have confidence percentages between 85.5% to 94.5% (so, 90% on average), there will be precisely one AI mistake. At the other end, in the set of stimuli pairs that have confidence percentages between 55.5% to 64.5% (i.e., 60% on average) there will be precisely four AI mistakes. As a result, of the 40 decisions, 30 will be correct, and thus the average accuracy of the Wizard-of-Oz AI model in the experiment is 75%.

Table 1

This matrix represents a set of 40 stimuli pairs, each with a unique AI confidence percentage and decision classification. Resulting in 10 incorrect AI decisions and 30 correct AI decisions.

Set \ Pair		1	2	3	4	5	6	7	8	9	10
85.5% - 94.5%	1 Mistake	False Negative	True Positive	True Positive	True Negative	True Positive	True Negative	True Positive	True Negative	True Positive	True Negative
75.5% - 84.5%	2 Mistakes	False Negative	False positive	True Positive	True Negative	True Positive	True Negative	True Positive	True Negative	True Positive	True Negative
65.5% - 74.5%	3 Mistakes	False Negative	False positive	False positive	True Negative	True Positive	True Negative	True Positive	True Negative	True Positive	True Negative
55.5% - 64.5%	4 Mistakes	False Negative	False positive	False positive	False Negative	True Positive	True Negative	True Positive	True Negative	True Positive	True Negative

Table 2

The decision classifications with respect to the experimental decisions "Same" vs "Different"

		Actual value	
		Positive (Same)	Negative (Different)
AI Decision	Positive (Same)	True Positive *The AI correctly decides "Same"*	False Positive *The AI incorrectly decides "Same"*
	Negative (Different)	False Negative *The AI incorrectly decides "Different"*	True Negative *The AI correctly decides "Different"*

4.5. Constructs and Measures

Objective measures. The study focuses on two primary objective measures to answer the hypotheses; accuracy and speed. Speed entails the average decision time per trial. Accuracy represents a percentage that represents the portion of correct trials (i.e., a percentage value between 0 and 100). These objective metrics are recorded during the experiment using Javascript and are finally exported to a .csv-file with the rest of the collected experimental data of the participant.

Subjective measures. Besides the objective measures, the study also includes three subjective measures. These subjective measures are measured immediately after the decision task

using multiple questionnaires with 5-point Likert scale response options. The first of the three subjective constructs is perceived workload. This construct is measured using the NASA Task Load Index (NASA-TLX), which is a widely used assessment tool that measures perceived workload using 6-items (Hart & Staveland, 1988). The second subjective construct is perceived time pressure, this construct has 4-items that have been formulated by myself. The third construct is system satisfaction. This construct is measured using the System Usability Scale (SUS-10) (Brooke, 1996). Each construct was found to have sufficient internal consistency (as discussed in further detail in the results). The questionnaires can be found in Appendix A.

4.6. Data Analysis

This section details the process from having the raw data .csv-files to executing the statistical analysis.

4.6.1. Data Preparation

Various steps are taken to prepare the data for statistical analysis. First the raw csv-files of the participants are merged into one dataset and the time units are converted from milliseconds to seconds for legibility. Then the data is inspected for outliers and errors. No errors are found. The outliers are detected by standardizing variables and detecting values that are deviating 3 standard deviations from the mean. No outliers on accuracy or speed are found. Six outliers on the duration of the experiment are found; while the experiment duration has a mean of 8.2 minutes and a median of 7.4 minutes, the six outliers took between 22 to 41 minutes to complete the experiment. Furthermore, five outliers on the duration of the task are found; while the task duration has a mean of 3.5 minutes and a median of 3.3 minutes, the five outliers took between 7.3 to 8.8 minutes to complete the task. These outliers (10 unique observations) are consequently

excluded from the analyses due to them not performing the experiment in the intended manner (i.e., without taking breaks), resulting in a total of 266 remaining observations.

4.6.2. Statistical Tests

Assumptions. As will be discussed in more detail below, the statistical analyses that are used are independent samples t-tests and one-way ANOVA tests. Therefore, various assumptions need to be checked.

Continuous Data. For both the t-test and the ANOVA analyses, an assumption is that the dependent variable is continuous. While speed (the average decision time for a trial) is continuous, the other objective measure, accuracy, is technically not continuous; the variable takes on values between 0 and 100 in steps of 2.5 (due to there being a total of 40 trials). Furthermore, the subjective measures are ordinal variables with 5 levels (resulting from the 5-point Likert scale questionnaires).

Normal distribution. Another assumption is that the dependent variable is normally distributed within each of the groups. Speed is transformed using the log function to make the distribution as normal as possible. As a result, speed is normally distributed in all conditions except condition 3. Accuracy is normally distributed in all conditions. Furthermore, all subjective measures except system satisfaction are normally distributed.

Homogeneity of variance. The assumption of homogeneity of variance is met in all groups on all dependent variables.

Independence of groups. The assumption of independent groups is also met as there is no dependence between the conditions.

Objective measures. The objective measures of interest are speed and accuracy. In order to test the general effect (on the objective measures) of the conditions and of the mere presence

of AI, an ANOVA test is performed including all four groups (i.e., including the "no-AI" baseline condition). When significant differences are found, a post-hoc analysis is performed to determine the group differences and their significance. Tukey's HSD is used to prevent *p*-value inflation due to multiple comparisons.

Continuing to the hypothesis testing, conceptually each of the hypotheses makes a comparison between two groups. Hypotheses H1a and H1b test the change in accuracy and speed between condition 3 and condition 2. Hypotheses H2a and H2b test the change in accuracy and speed between condition 4 and condition 2. For this reason, four separate independent samples t-tests are used to answer the hypotheses H1a, H1b, H2a, and H2b.

Subjective measures. First, the reliability of the three constructs of the three subjective measures (workload, time pressure, system satisfaction) are tested using Cronbach's alpha. Then for each of the subjective measures an ANOVA analysis is conducted on the four groups to test for differences between conditions. When significant differences are found, this is followed by a post-hoc analysis that includes the Tukey correction for multiple comparisons.

5. Results

This section presents the results of the analyses on the data from the experiment. The results of the objective measures are presented first, followed by the results of the subjective measures. Each section first discloses descriptive statistics prior to the inferential test results.

5.1. Objective Measures

Tables 3 and 4 report the descriptive statistics including the mean and standard deviation of accuracy and average decision time. As the tables show, averaged across all conditions, the accuracy is on average 59.8% and the average decision time is 3.7 seconds.

Table 3
Descriptive Statistics of Accuracy (% of Trials Correct)

Condition	Mean	Std. Dev.	N
1	58.4	8.6	63
2	59.7	8.6	67
3	60.1	9.7	69
4	60.7	9.3	67
Total	59.8	9.1	266

Table 4
Descriptive Statistics of Average Decision Time (Average Decision Time in Seconds)

Condition	Mean	Std. Dev.	N
1	3.3	1.0	63
2	3.6	1.0	67
3	4.0	1.2	69
4	3.9	1.2	67
Total	3.7	1.1	266

5.1.1. Accuracy

The means and distributions of accuracy within each condition are represented in Figure 6 and 7, respectively. An ANOVA analysis is performed with condition as the independent variable and accuracy as the dependent variable. The test indicates no significant differences between the groups ($F = .71$, $p = .55$, R-squared = .01).

Furthermore, two independent samples t-tests are performed to compare accuracy between condition 2 and 3 (hypothesis 1a) and between condition 2 and 4 (hypothesis 2a). No significant differences are found in either of the t-tests (respectively, $p = .76$, $p = .51$), meaning that there are no significant differences in accuracy between conditions 2 and 3, nor between conditions 2 and 4. Hence, hypotheses 1a and 2a are not supported by the results.

Figure 6
Means and Confidence Intervals of Accuracy Within Each Condition

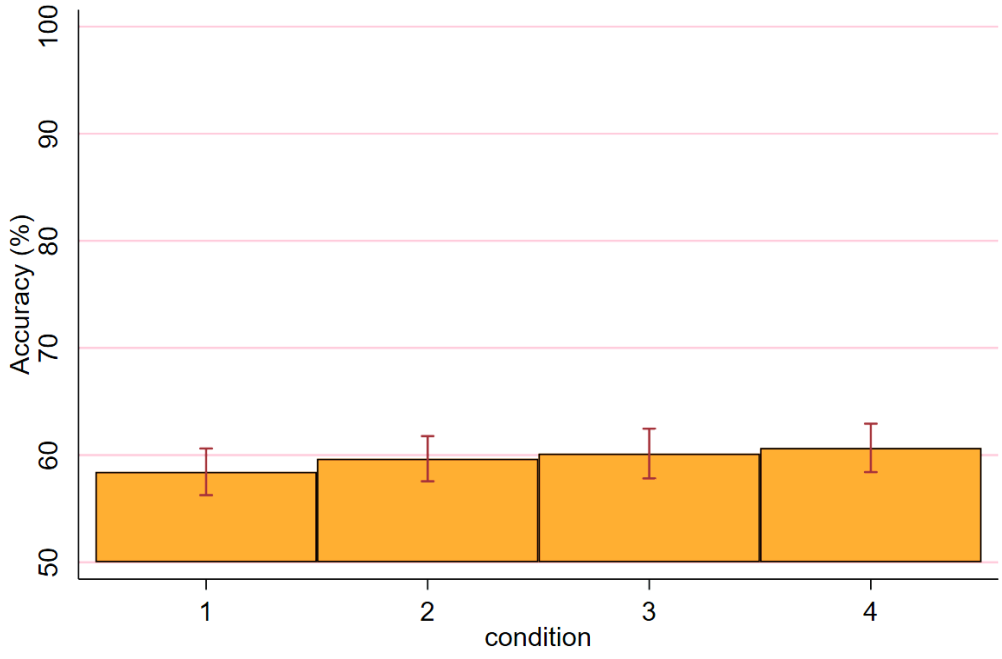
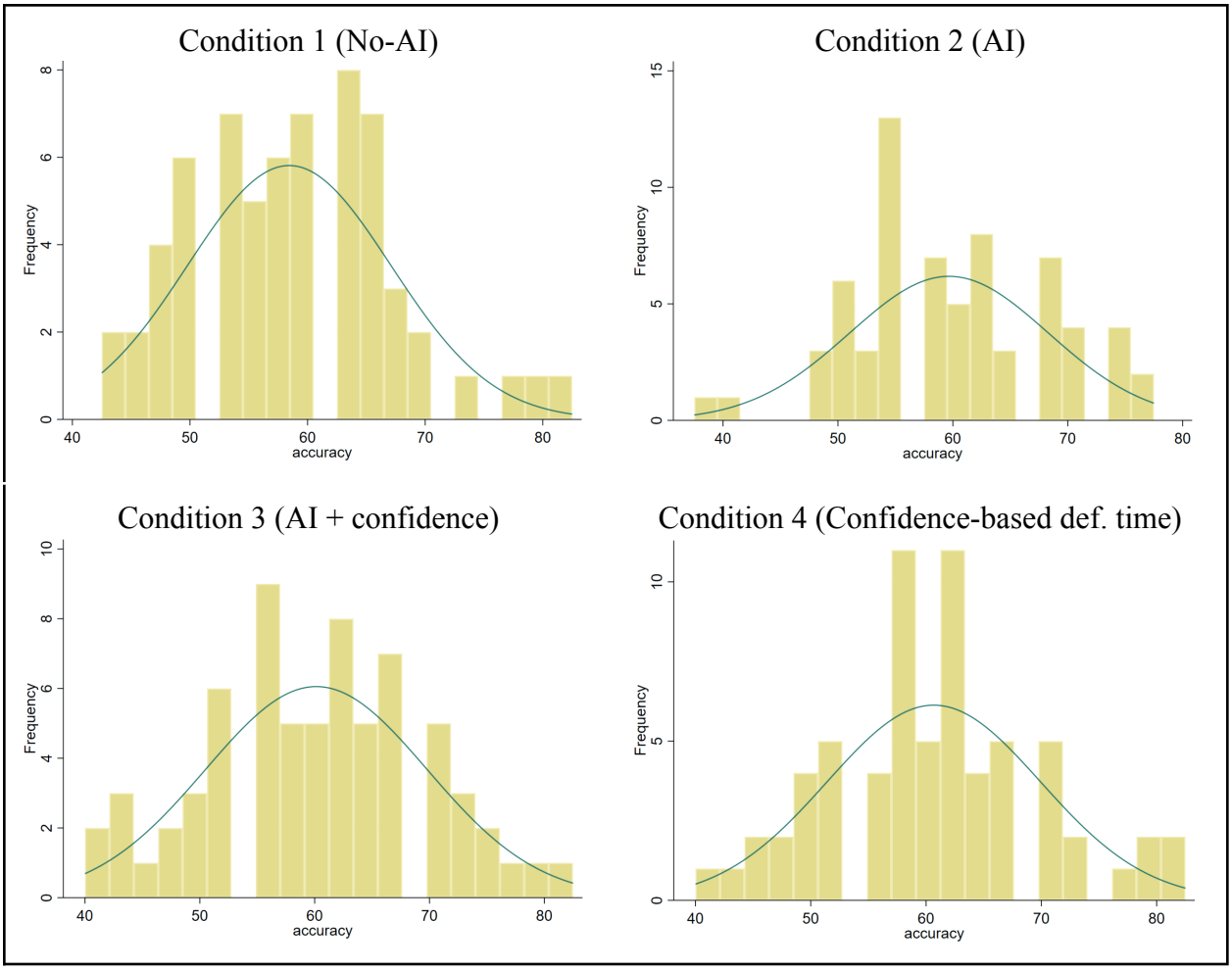


Figure 7
Distributions of Accuracy Within Each Condition



5.1.2. Speed

The average decision time variable is transformed to the logarithmic variant to meet the normal distribution assumption. The means and distributions of this variable within the conditions are represented in Figure 8 and 9, respectively. An ANOVA analysis is performed with condition as the independent variable and the logarithmized average decision time as the dependent variable. The test indicates there are significant differences between the groups (F = 6.92, $p < .01$, R-squared = .07). A post-hoc analysis reveals significant differences between condition 1 and condition 3 ($p < .01$) and between condition 1 and condition 4 ($p < .01$). The

contrast values are, respectively, .21 and .18. For the non-transformed variable of average decision time there is, respectively, a .75 and a .62 difference in the means between the conditions. Meaning that in conditions 3, participants took on average .75 seconds longer to make a decision compared to participants in condition 1, and in condition 4, participants took on average .62 seconds longer to make a decision compared to condition 1.

Two independent samples t-tests are performed on the logarithmized average decision time to compare the decision speed between condition 2 and 3 (hypothesis 1b) and between condition 2 and 4 (hypothesis 2b). The H1b t-test finds a significant *p*-value of .02 for the difference in means between condition 2 and condition 3, where condition 3 has a .11 higher mean value of logarithmized average decision time than condition 2. Doing the t-test with the non-transformed variable of average decision time similarly finds a p-value of .02 and finds that condition 3 has a .44 seconds higher mean value for average decision time than condition 2. Hence, hypothesis 1b is supported by the results; the participants were significantly slower in condition 3 than in condition 2, namely, on average they took an additional .44 seconds. Calculating the effect size of this significant effect results in a Cohen's d of .40, which is considered a medium effect. The t-test of hypothesis 2b finds no significant results ($p = .1$). Hence, hypothesis 2b is not supported by the results.

Figure 8

Means and Confidence Intervals of Average Decision Time (Bottom One is Logarithmized)

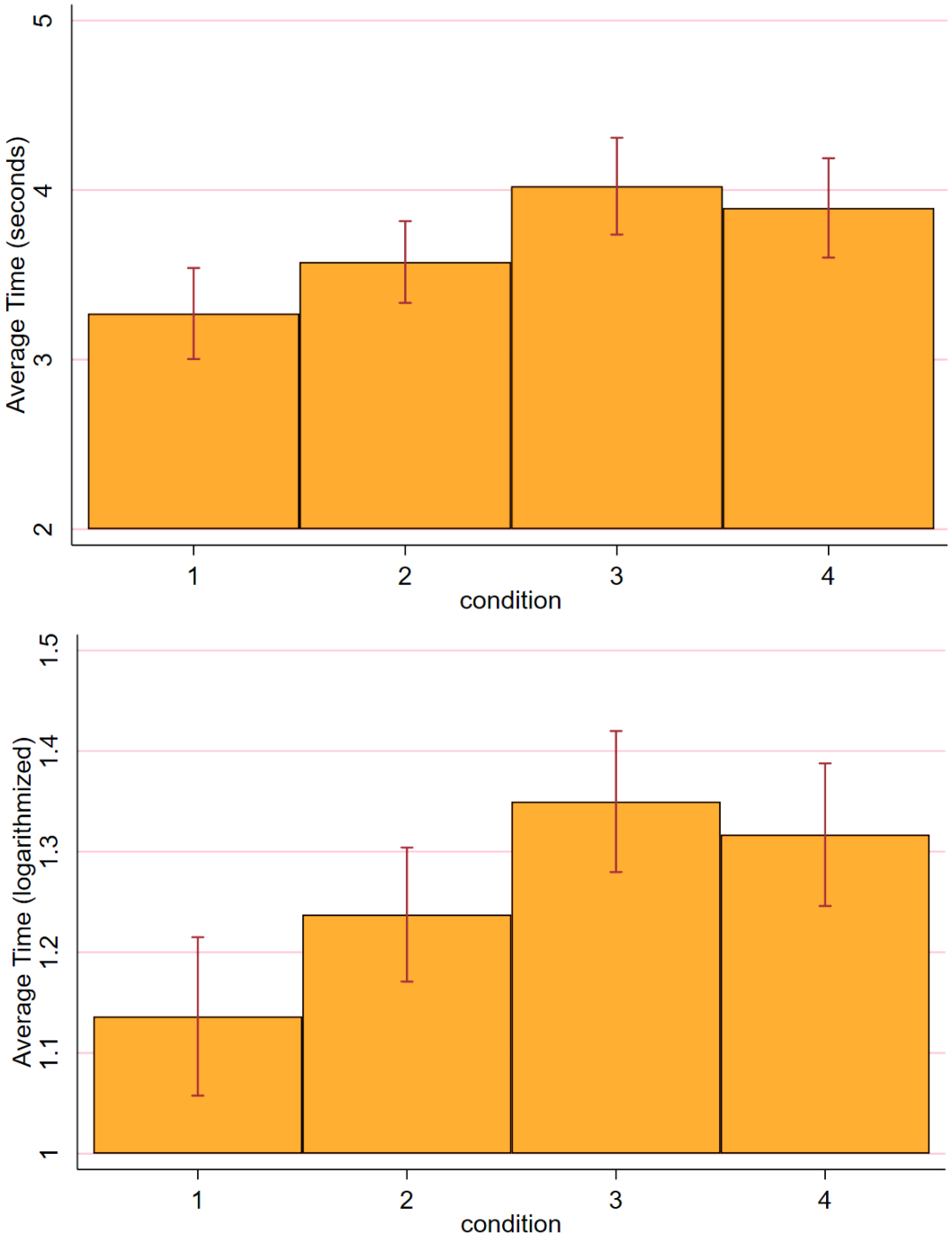
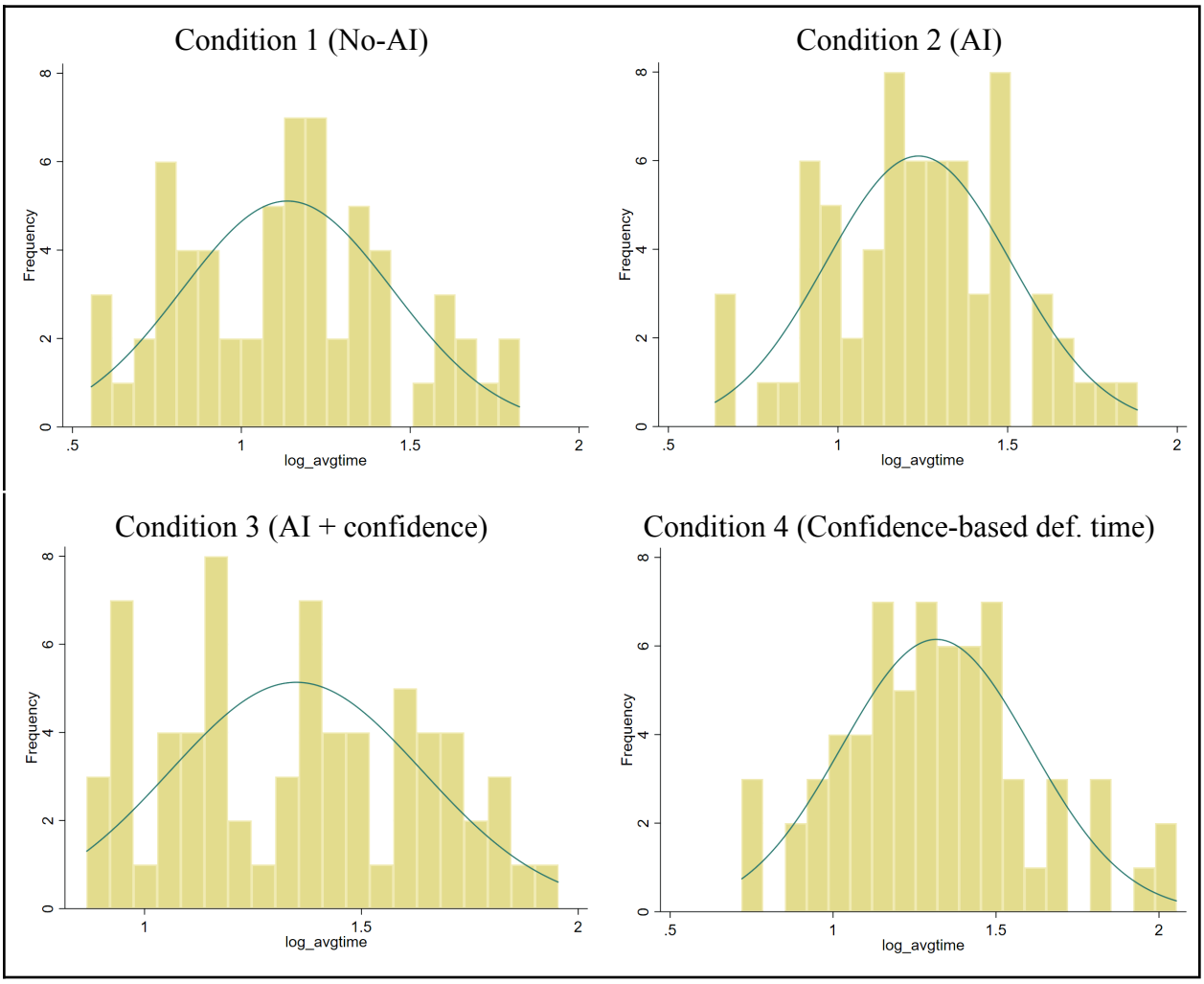


Figure 9

Distributions of the Logarithmized Average Decision Time Within Each Condition



5.2. Subjective Measures

Table 5, 6, and 7, report the descriptive statistics including the mean and standard deviation of perceived workload, time pressure, and system satisfaction, respectively. The constructs can theoretically take values between 0 and 4, with 4 representing maximal workload, time pressure, or satisfaction, and 0 representing minimal. The descriptive tables show that the no-AI condition has the lowest perceived workload and highest perceived system satisfaction. Furthermore, that condition 3 scores the highest perceived workload, lowest system satisfaction, and highest time pressure.

Table 5 shows that overall the perceived workload was rated relatively low with 1.6 out of 4. Across the conditions, participants in condition 1 perceived the least workload while participants in conditions 3 and 4 perceived the most.

Table 5
Descriptive Statistics of Perceived Workload

Condition	Mean	Std. Dev.	Min	Max	N
1	1.4	.6	.2	3.2	63
2	1.5	.7	0	3.3	67
3	1.7	.6	.5	2.8	69
4	1.7	.7	.2	3.3	67
Total	1.6	.6	0	3.3	266

Table 6 shows that overall the time pressure was perceived as moderate at 1.8 out of 4. Across the conditions, participants in condition 3 perceived the most time pressure.

Table 6
Descriptive Statistics of Time Pressure

Condition	Mean	Std. Dev.	Min	Max	N
1	1.9	.8	0	4.0	63
2	1.7	.8	0	3.8	67
3	2.0	.9	0	4.0	69
4	1.8	.9	0	3.5	67
Total	1.8	.9	0	4.0	266

Finally, Table 7 shows that overall the satisfaction with the system was rated relatively high with an overall score of 3.1 out of 4. Across the conditions, participants in condition 3 perceived the system as the least satisfying.

Table 7
Descriptive Statistics of System Satisfaction

Condition	Mean	Std. Dev.	Min	Max	N
1	3.2	.4	1.9	4.0	63
2	3.2	.5	2.0	4.0	67
3	3.0	.5	1.9	4.0	69
4	3.1	.5	1.9	3.9	67
Total	3.1	.5	1.9	4.0	266

5.2.1. Perceived Workload

The items of the NASA-TLX are used to construct the variable perceived workload. The construct has a Cronbach's alpha of .70, which is considered sufficiently acceptable internal consistency. The means and distributions of the variable within the conditions are represented in Figure 10 and 11. An ANOVA analysis is performed using the construct as the dependent variable and condition as the independent variable. The ANOVA test indicates that there are no statistically significant differences between the conditions. ($F = 2.40$, $p = .07$, R-squared = .03).

Figure 10

Means and Confidence Intervals of Perceived Workload Within Each Condition

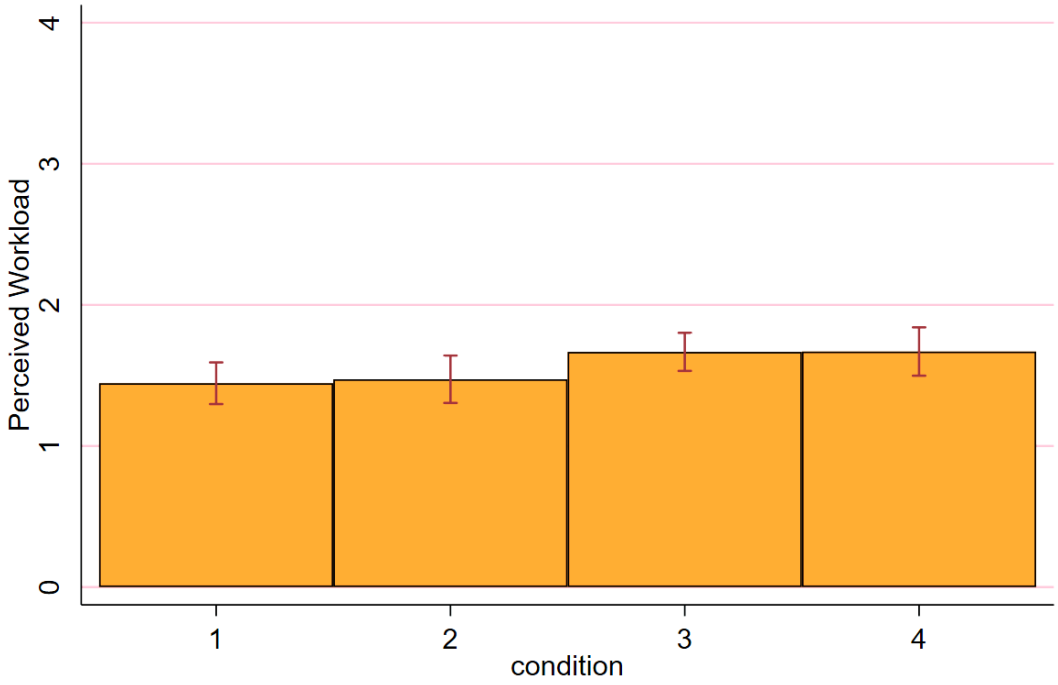
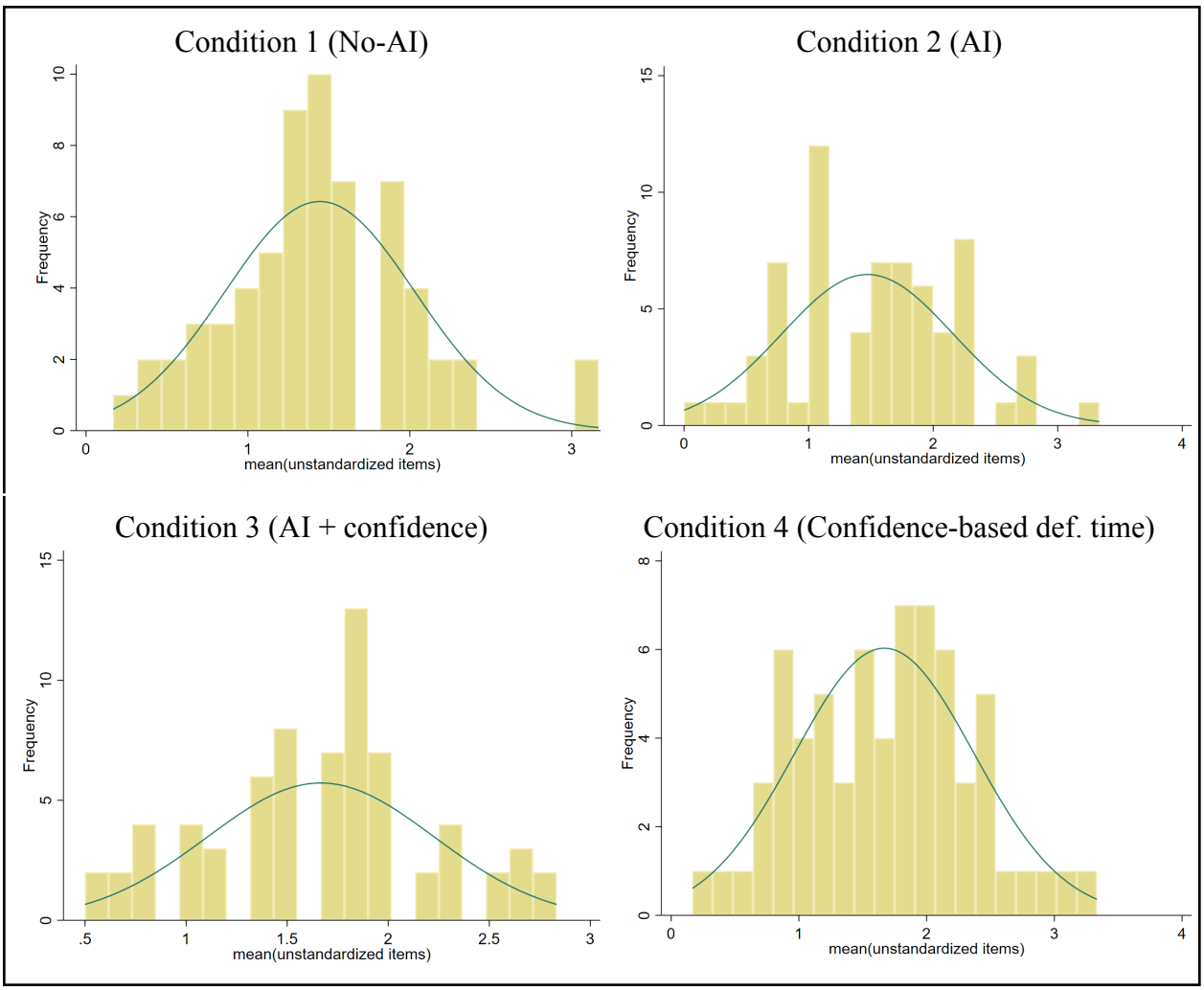


Figure 11

Distribution of Perceived Workload Within Each Condition



5.2.2. Time Pressure

Four items have been formulated to represent time pressure. The construct has a Cronbach's alpha of .86, which indicates good internal consistency. The means and distributions of the variable within the conditions are represented in Figure 12 and 13, respectively. An ANOVA analysis is performed using the standardized construct as the dependent variable and condition as the independent variable. The ANOVA test indicates that there are no statistically significant differences between the conditions ($F = 1.33$, $p = .26$, R-squared = .02).

Figure 12
Means and Confidence Intervals of Time Pressure Within Each Condition

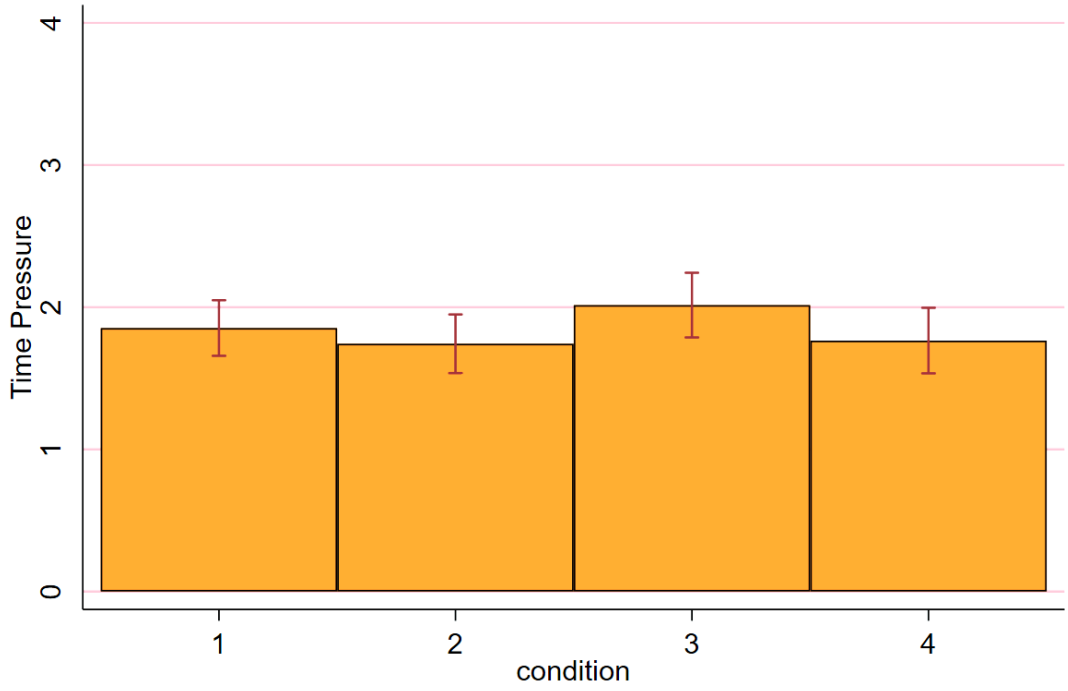
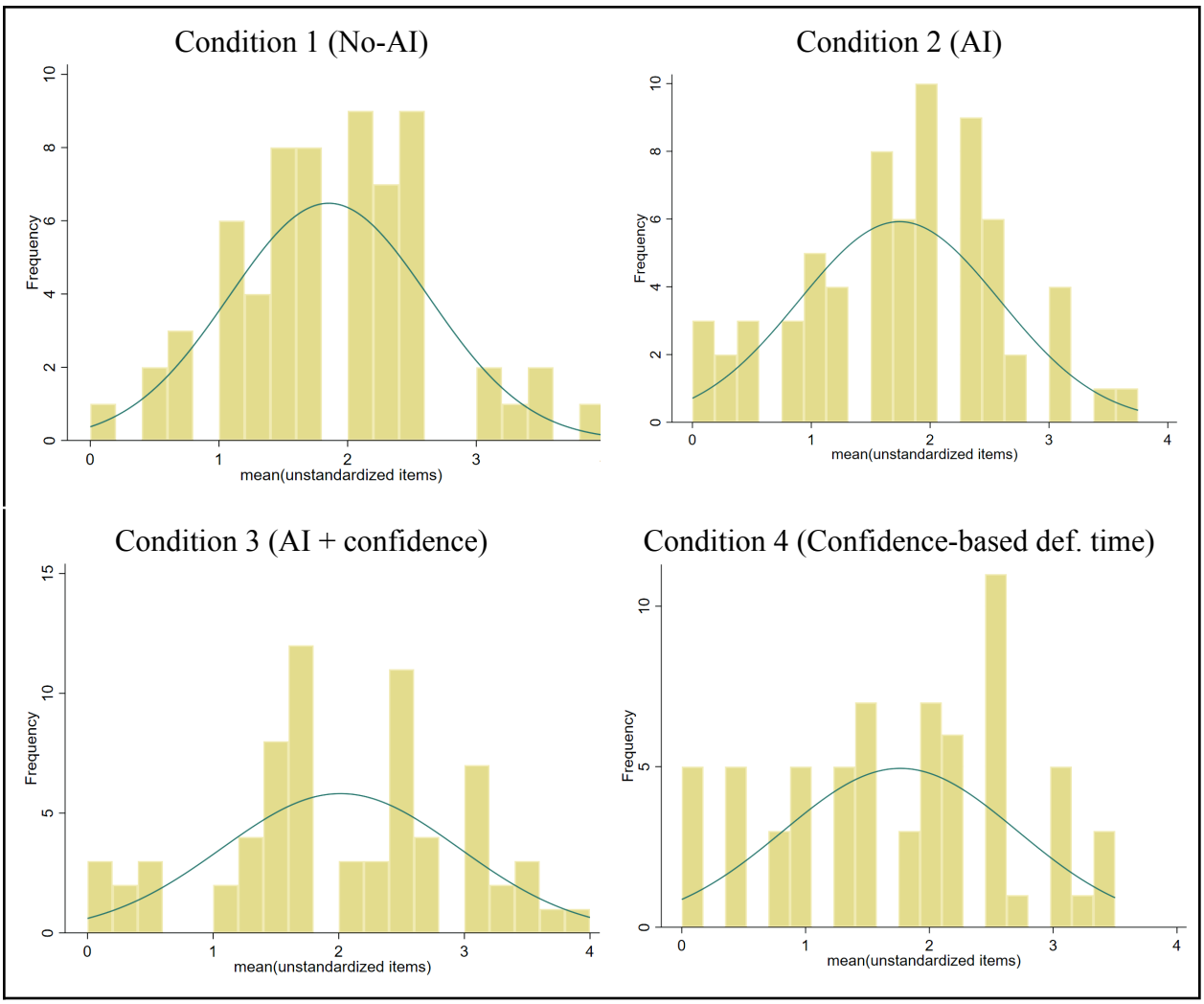


Figure 13

Distribution of Time Pressure Within Each Condition



5.2.3. System Satisfaction

The items of the SUS-10 are used to generate a construct for system satisfaction. The construct has a Cronbach's alpha of .79, which indicates good internal consistency. The means and distributions of the variable within the conditions are represented in Figure 14 and 15. An ANOVA analysis is performed using the construct as the dependent variable and condition as the independent variable. The ANOVA test indicates that there are no statistically significant differences between the conditions ($F = 1.41$, $p = .24$, R-squared = .02).

Figure 14
Means and Confidence Intervals of System Satisfaction Within Each Condition

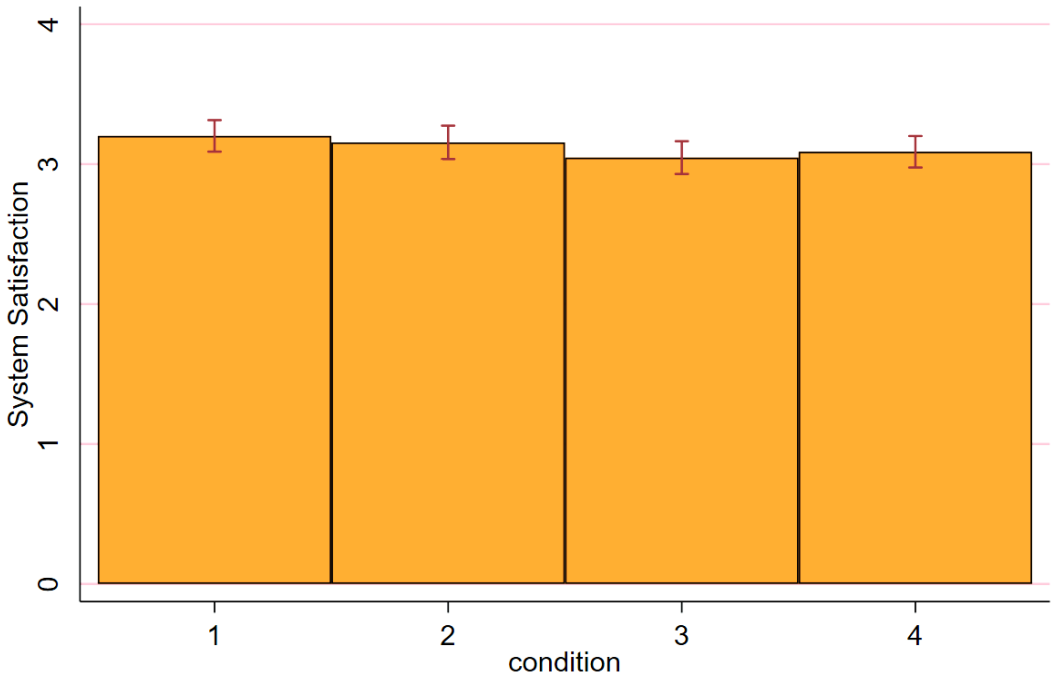
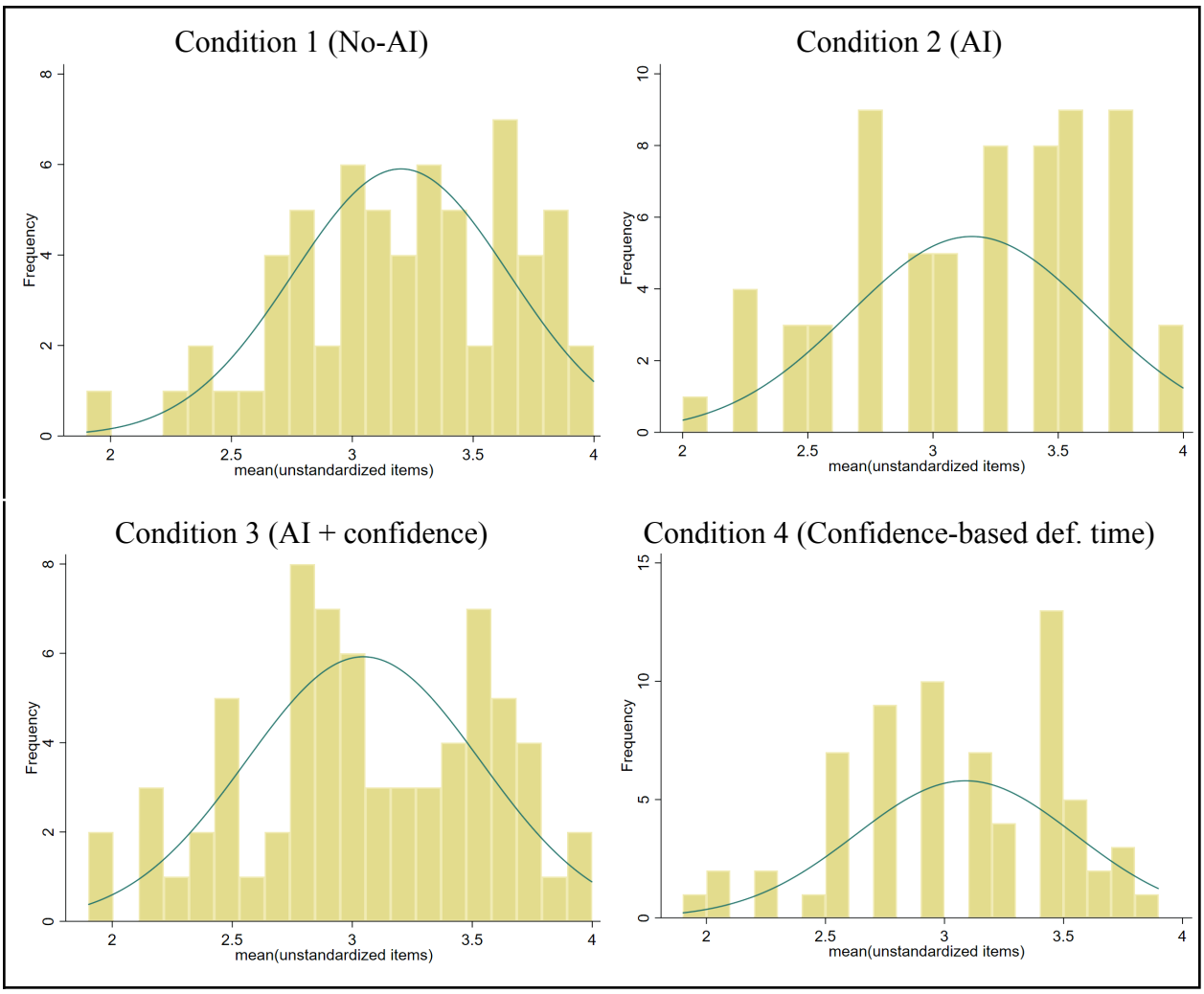


Figure 15

Distribution of System Satisfaction Within Each Condition



6. Discussion

In this section, the results will be interpreted and contextualized by comparing it with previous findings in the literature. Furthermore, nuance to the results will be introduced as the limitations of the study are discussed. Finally, the practical and theoretical implications of the findings of the study are revealed and suggestions for future research are provided.

6.1. Interpretation of the Results

The findings find no support for an increase in joint human-AI decision-making accuracy with the presentation of confidence information (H1a). Neither do the findings support an increase in decision-making accuracy when the deferral time is based on the AI's decision's confidence (H2a). Furthermore, the findings do not find support that a confidence-based deferral time leads to an increase in joint human-AI decision-making speed (H2b). The findings do find support for the hypothesis that the presentation of confidence information slows down the decision-making speed (H1b). This finding may be attributed to the increase in cognitive load that results from being presented with the confidence information. Namely, the confidence information presents extra information that requires processing on the participants' behalf. However, the measurement of perceived workload did not find a significant difference between the conditions. The other two subjective measurements; time pressure and system satisfaction, also did not differ significantly between any of the conditions.

In the decision task, the AI model has an accuracy of 75%. The participants in the baseline condition (no AI) have an accuracy of 58.4%, which indicates that the task is quite difficult for the participants. It is interesting that in the conditions with AI assistance, the accuracy scores remain close to the baseline figure (namely, 59.7%, 60.1%, and 60.7%). This suggests that the participants, irrespective of the presence of the AI, tend to follow their own

judgements. To put the accuracy difference in terms of trials, without AI assistance, participants have about 23 out of 40 trials correct, and with AI assistance, they have about 24 out of 40 trials correct. The rather small difference may be due to a variety of reasons:

Firstly, it could be that the participants are overestimating their own ability for the task. During the experiment, participants do not receive any feedback on their decisions, which doesn't allow them to observe their own aptitude at the task. When participants have no reason to suspect their judgment may be suboptimal, they may only attend to the AI's decision when they are unsure. Research by Bondi et al. (2021) finds that people are much less likely to follow the model's predictions when their confidence in their own aptitude is high. Furthermore, research by Chong et al. (2022) finds that it is human self-confidence—not their confidence in the AI—that directs their decision to accept or reject the AI's suggestions.

Secondly, it could be that the participants did not trust the AI sufficiently. There is no possibility for the participant to gauge the quality or trustworthiness of the AI model as they receive no feedback on the decisions. All they can do is rely on the task description, which states that the model on average has a 75% accuracy.

Another potential cause of the small difference in accuracy between the no-AI and AI conditions, is that participants may enjoy doing the task more by themselves. This could be because they simply find the facematching challenge enjoyable or because the AI assistance increases the cognitive load. Simply going with their own judgment and not having to consider information that the AI provides is a strategy that requires less effort. Arguably, the least-effort method is to blindly follow the AI's judgements on each decision, however, doing this would appear obvious to the experiment leader and does not follow the protocol of the experiment.

6.2. Contextualization of the Results

This section contextualizes the results by viewing them in light of earlier findings in the joint human-AI decision-making literature.

The finding that the presentation of confidence does not lead to an increase in accuracy aligns with various findings in earlier literature. These findings suggested that, although AI confidence disclosure may increase trust in the system, it does not lead to an increase in trust calibration which would be needed for accurate decision-making (Buçinca, 2021; Lai, 2019). Furthermore, as asserted by Zhang et al. (2020), even when trust is properly calibrated due to confidence presentation, an improvement on decision accuracy would still not show if the error boundaries of the human and the AI are largely the same and they are unable to complement each other. It should be noted, however, that the impact of confidence presentation on trust in the AI has not been established in this study. Therefore, it remains uncertain whether confidence presentation had any impact on trust in the AI at all.

When it comes to the speed of joint human-AI decision-making, a significant finding of this study is that presenting confidence levels results in slower decision-making. This finding is consistent with the cognitive understanding of decision-making, which suggests that processing a larger amount of evidence requires more cognitive processing, leading to a longer decision time on average.

Research by Buçinca et al (2021), suggested that people rate the less cognitively forcing conditions more favorably (although these conditions were performing worse). However, this study did not find significant differences in system satisfaction between the conditions. Furthermore, research by Buçinca et al. (2021) found that people often over rely on AI systems and accept their suggestions even when they are incorrect. However, in this study, based on the

accuracy scores in each of the conditions, it appears that people in the AI conditions under relied on the AI.

With respect to the lack of statistical support for the benefits of a confidence-based deferral time, there have been no prior studies investigating exactly this approach. It is evident that more research will be needed into this deferral approach and no definitive conclusions can be drawn yet about the effectiveness of such an approach.

6.3. Limitations

The experiment is subjected to a number of limitations which are discussed in detail below.

Lack of certain measures. The study does not include a measure of the participant's trust in the AI, trust calibration, trust in oneself, or a measure to track the desire or tendency of the participant to collaborate with the AI. The results suggest that the participants' trust or collaboration efforts with the AI are remarkably low and that as a result the AI manipulations did not have as much effect as expected. The data to confirm this suspicion is lacking, however, as the required measures to confirm this have not been included in the study.

Context specificity. The study uses a specific decision task; face-matching. As a consequence, the findings lack generalizability to other types of decision tasks. Furthermore, the study uses a model with an accuracy of 75% and a decision task with human accuracy of roughly 60%. There is nothing wrong with these numbers per se, but they do measure a specific context and therefore the results may not translate to models and decision tasks with different accuracy levels. Lastly, the fixed deferral time is set at 13 seconds and the confidence-based deferral time ranged between 9 and 17 seconds. The number of these parameters influence the effects of the study. As a result, the findings of this study may not necessarily generalize to contexts with different deferral times.

Celebrity stimuli. The stimuli features celebrities; some of the participants may be familiar with one or more of the celebrities featured in the data set and as a result may have an improved performance on those trials, consequently impacting the accuracy and speed of those particular trials.

Online recruitment and online experiment. The participants were recruited online through a participant-recruitment-platform. These participants typically participate in multiple experiments a day as a way to earn money. It is therefore unclear how incentivized these participants are to perform the experiment to the best of their ability. Furthermore, the lack of a lab setting makes it impossible to control extraneous variables such as interruptions and other environmental factors that affect the experiment. As a result, there is reduced ecological validity to the results of the study.

6.4. Implications

The results of this study have both practical and theoretical implications. Firstly, the findings suggest that incorporating confidence information may not necessarily lead to improved joint human-AI decision-making accuracy and may actually slow down the decision-making. This is important information to consider for industries and organizations that are looking to implement AI in decision-making processes, as they may need to reconsider whether or not to present AI confidence information to human decision-makers. For instance, a slowdown of joint human-AI decision-making time may have severe implications for domains dealing with a huge number of incoming decisions that need to be handled in a timely manner, such as security screening or content moderation. For content moderation, for instance, it is important that harmful content does not linger on the platform any longer than it has to. In conclusion, designers of such joint human-AI decision-making systems should be critical about the

information they present to the human decision-maker during decision deferral, and calibrate it in such a way that optimizes both metrics, accuracy and speed, for their specific use case. Furthermore, the finding that participants tend to follow their own judgements, even in the presence of the AI model, underlines the importance of designing human-AI decision-making systems that promote collaboration between the human and the AI, consequently improving the performance of such collaborative systems. This is another practical implication that designers of such systems should take into account—to design the system in a way that the AI's capabilities are properly utilized by the human decision-maker.

Another practical implication is that, at present, there is no established scientific foundation for organizations to implement a confidence-based deferral time in their human-AI decision-making processes, as its potential benefits have not yet been established. This suggests that, currently, other joint human-AI decision-making approaches may be more effective. Hence, a theoretical implication of the results is that there is a need for more research into the effects of the implementation of a confidence-based deferral time. Furthermore, another theoretical implication is the need for more research into the effects of confidence presentation. The need for further research is discussed in more detail in the following section.

6.5. Future Research

With the concept of a confidence-based deferral time, this research provides a new approach to joint human-AI decision-making. In light of this study being one of the first to research this concept, and having a number of limitations, more research will be required to make conclusive statements about the potential benefits of this approach. Future research on this approach should include a variety of decision tasks, as the approach may be effective in certain types of decision tasks and ineffective in others. For instance, one could research the effect the

complexity of the task has on the effectiveness of this approach. Or the duration of the decision task. By researching the approach in different contexts, it will add more nuance and detail to the scientific understanding of the approach. Furthermore, findings become more generalizable and conclusive when they emerge across a range of different contexts. Future research could also consider using larger samples and longer decision tasks, as this increases the power and the reliability of the findings. Furthermore, as the AI used in this study has an accuracy of 75%, and the humans on their own had an accuracy of roughly 60%, future research could consider using AI models with different levels of accuracy and could consider decision tasks with different levels of human accuracy, to see how this affects the joint human-AI decision-making process and in particular the effectiveness of a confidence-based deferral time. Future research should ensure participants are incentivized to perform to the best of their ability and hence should consider linking performance to the payment. Future research should also explore what happens when participants receive feedback in-between the decisions, and how this impacts their trust in the AI or in themselves, and the consequential effect on the decision-making performance under a confidence-based deferral time. Continuing on the notion of trust, future research should also aim to explore the importance of trust calibration with respect to the efficiency of a confidence-based deferral time. As was discussed in the implications, without proper trust (calibration), the effects of AI manipulations will not properly emerge as the AI's judgment will be largely dismissed by the participants. Finally, future research should also be conducted to explore different methods of presenting confidence information and the impacts these methods have on decision accuracy and speed, as the current best practice regarding the inclusion of confidence information and the manner, remains unclear.

7. Conclusion

The findings of the study show that showing confidence information slows down the speed of joint human-AI decision-making while not necessarily leading to an improvement in decision accuracy. Furthermore, the results of the study provided no statistical support for the effectiveness of a confidence-based deferral time, as neither the accuracy nor the speed of the decision-making improved in the joint human-AI decision-making task. As one of the first studies to test this new approach, this study provides a valuable contribution to the field of AI and decision-making, and highlights the need for further research in this area to increase our understanding of how to incorporate AI prediction confidence in joint human-AI decision-making processes and to improve our understanding of the potential benefits of a confidence-based deferral time.

References

Allen, J. E., Guinn, C. I., & Horvtz, E. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications, 14*(5), 14-23.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., & Weld, D. (2021, May). 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* (pp. 1-16). https://doi.org/10.1145/3411764.3445717

Bondi, E., Koster, R., Sheahan, H., Chadwick, M., Bachrach, Y., Cemgil, T., Paquet, U., & Dvijotham, K. (2022, June). Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 5, pp. 5286-5294). https://doi.org/10.1609/aaai.v36i5.20465

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). 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), 1-21. https://doi.org/10.48550/arXiv.2102.09692

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research, 70*, 245-317. https://doi.org/10.1613/jair.1.12228

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics* (pp. 160-169). IEEE. https://doi.org/10.1109/ICHI.2015.26

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Coxon, M. (2012). *Cognitive psychology*. Learning Matters.

Cveticanin, N. (2023, January 20). Unfiltered Instagram Statistics You'll Want to Share With All Your Followers. *Dataprot*. https://dataprot.net/statistics/instagram-statistics/

Dambacher, M., & Hübner, R. (2015). Time pressure affects the efficiency of perceptual processing in decisions under conflict. *Psychological research*, *79*, 83-94. https://doi.org/10.1007/s00426-014-0542-z

De Ridder, D., Kroese, F., Adriaanse, M., & Evers, C. (2014). Always gamble on an empty stomach: Hunger is associated with advantageous decision making. PloS one, 9(10), e111081. https://doi.org/10.1371/journal.pone.0111081

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2021). The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354.* https://doi.org/10.48550/arXiv.2105.03354

Donkin, C., Little, D. R., & Houpt, J. W. (2014). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 1183. https://doi.org/10.1037/a0035947

Face Comparison / Face Similarity Test Online. (n.d.). ToolPie. https://facecomparison.toolpie.com/

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kobie, N. (2018, October 18). Heathrow's facial recognition tech could make airports more bearable. *WIRED UK*. https://www.wired.co.uk/article/heathrow-airport-facial-recognition-technology

Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in artificial intelligence*, *4*, 622364. https://doi.org/10.3389/frai.2021.622364

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*. https://doi.org/10.48550/arXiv.2112.11471

Lai, V., & Tan, C. (2019, January). 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* (pp. 29-38). https://doi.org/10.1145/3287560.3287590

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., & Bermúdez i Badia, S. (2021, May). A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-14). https://doi.org/10.1145/3411764.3445472

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. *Psychological science, 25*(9), 1663-1673. https://doi.org/10.1177/0956797614541991

Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, *31*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence, 267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 427-436).

Petrov, C. (2023, January 12). 50+ Stunning Twitter Statistics You Need to Know in 2023. *Techjury*. https://techjury.net/blog/twitter-statistics/

Prolific · Quickly find research participants you can trust. (n.d.). https://www.prolific.co/

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1226–1243. https://doi.org/10.1037/a0036801

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., & Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports, 12*(1), 14952. https://doi.org/10.1038/s41598-022-18751-2

Siegel, A., & Sapru, H. N. (2005). Essential neuroscience. Philedelphia.

Tegmark, M. (2018). Life 3.0: *Being human in the age of artificial intelligence*. Vintage.

TrueList. (2023, January 7). Facebook Statistics 2023. *TrueList*. https://truelist.co/blog/facebook-statistics/

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., & Li, Y. (2022, June). Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning* (pp. 23631-23644). PMLR. https://doi.org/10.48550/arXiv.2205.09310

Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, *292*(1), 60-66. https://doi.org/10.1148/radiol.2019182716

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019, May). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1-12). https://doi.org/10.1145/3290605.3300509

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 295-305). https://doi.org/10.1145/3351095.3372852

Appendix A

Perceived Workload (NASA-TLX)

Please indicate your answer by selecting one of the empty boxes:

	1 (Low)	2	3	4	5 (High)
Mental Demand How mentally demanding was the decision-making task?					
Physical Demand How physically demanding was the decision-making task?					
Temporal Demand How hurried or rushed was the pace of the decision-making task?					
Performance How successful do you feel you were in accomplishing what you were asked to do?					
Effort How hard did you have to work to accomplish your level of performance?					
Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?					

Time Pressure

Please indicate to what extent you agree with the statements by selecting one of the empty boxes:

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
With respect to the timer, I think there was enough time to complete the decision-making task.					
Due to the timer, I felt in a hurry during the decision-making task.					
Due to the timer, I felt time pressure during the decision-making task.					
The timer provided me with enough time to make the decisions.					

System Satisfaction

Please indicate to what extent you agree with the statements by selecting one of the empty boxes:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I think that I would like to use this system frequently.					
I found the system unnecessarily complex.					
I thought the system was easy to use.					
I think that I would need the support of a technical person to be able to use this system.					
I found the various functions in this system were well integrated.					
I thought there was too much inconsistency in this system.					
I would imagine that most people would learn to use this system very quickly.					
I found the system very cumbersome to use.					
I felt very confident using the system.					
I needed to learn a lot of things before I could get going with this system.					

Bottom of questionnaire

If you have encountered any issues during the experiment or have any remarks, please describe them below:

Type here for issues/remarks...

Click to finish the experiment