Cover Page

THE EFFECT OF PERCEIVED TASK DIFFICULTY ON RELIANCE ON AUTOMATED FACIAL RECOGNITION SYSTEMS AS DECISION AIDS

[redacted]

This thesis is submitted in partial fulfilment of the Honours degree of Bachelor of Psychological Science (Honours)

Word count: 6966

Contents

List of Figures

Abstract

Automated Facial Recognition Systems (AFRS) are being used to verify an individual's identity by governments, businesses, and passport officers, while human operators monitor the system. There have been accidents reported when humans and automation work together, primarily due to under-trusting or over-trusting of the system. Thus, instilling an appropriate level of trust, and in turn, reliance on the system is important. A factor that has been found to influence trust in automation is perceived task difficulty. However, no research has been done to check whether perceived task difficulty influences reliance in AFRS, which we examined in our study. A total of 48 participants were administered Glasgow Face Matching Test 2-S and were randomly allocated into one of two conditions: the 'Easy' condition, who were told that the task would be easy, and the 'Hard' condition who were told the task would be hard. They completed the task with the assistance of the AFRS. On analysis, we did not find any difference between reliance on the AFRS in the two difficulty conditions. However, the participants did improve their accuracy with the use of AFRS, but there was no difference in improvement in the accuracy scores between the two conditions. No significant relationship was found between self-assessment of task difficulty and reliance on the AFRS. This raises important questions about whether the perception of task difficulty influences reliance on AFRS, which in turn will have important implications on the use of AFRS as a decision aid.

Keywords: Automated Facial Recognition System, Perceived Task Difficulty, Reliance, Trust

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University, and, to the best of my knowledge, this thesis contains no material previously published except where due reference is made. I give permission for the digital version of this thesis to be made available on the web, via the University of Adelaide's digital thesis repository, the Library Search and through web search engines, unless permission has been granted by the School to restrict access for a period of time.

Contributor Roles Table

ROLE	ROLE DESCRIPTION	STUDENT	SUPERVISOR 1	SUPERVISOR 2
CONCEPTUALIZATION	Ideas; formulation or evolution of overarching research goals and aims.	x	x	
METHODOLOGY	Development or design of methodology; creation of models.	x	x	
PROJECT ADMINISTRATION	Management and coordination responsibility for the research activity planning and execution.	x	x	x
SUPERVISION	Oversight and leadership responsibility for the research activity planning and execution, including		x	x

	mentorship external to the core team.			
RESOURCES	Provision of study materials, laboratory samples, instrumentation, computing resources, or other analysis tools.		x	
SOFTWARE	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code.		x	
INVESTIGATION	Conducting research - specifically performing experiments, or data/evidence collection.	x	x	

VALIDATION	Verification of the overall replication/reproducibility of results/experiments.	x	x	
DATA CURATION	Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.		x	
FORMAL ANALYSIS	Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data.	x		

VISUALIZATION	Visualization/data presentation of the results.	x		
WRITING – ORIGINAL DRAFT	Specifically writing the initial draft.	x		
WRITING – REVIEW & EDITING	Critical review, commentary, or revision of original draft	x	x	x

Introduction

Today, the world of technology continues to advance at a rapid pace. Many of the roles that humans were once solely responsible for are now being done with the aid of automation. Whether it be the defence and security sector, navigation, or border control, automation continues to reduce human workload. Nevertheless, humans still play a role in monitoring and checking the decisions. For example, autopilot is commonly used for guiding an airplane, however, a pilot is still needed to monitor the flight.

Automated Facial Recognition Systems (AFRS)

With advancements in technology, there are several ways to verify the identity of an individual using biometrics, ranging from thumbprints to eye scans to faces. Automated Facial Recognition Systems (AFRS) are commonly used by governments, security personnel, passport officers and private businesses for facial identification. They are an efficient and quick method of face matching. A common example of an AFRS is the electronic passport gates, also called e-Gates, that are used at international airports (Noyes & Hill, 2021). This is called a 1:1 image matching where a person's live-capture image is compared to their passport photo. In this process, the AFRS algorithm first finds the face in the two images and processes the relevant features in them. A similarity score is then calculated, which is used to detect whether the two images show the same face. A *high* similarity score suggests that the two images of the faces are the same, whereas a *low* similarity score suggests that the pictures are of two different people. These systems can also be used to compare an image against a database of images, for example, an image of a suspect captured by CCTV being compared against the images in the police database. The accuracy of early AFRS algorithms was limited as they could not perform well on pictures that differed in terms of poses, expressions, or lighting (Sengupta et al., 2016). However, the newer algorithms can detect similarities across a range of pictures (Taigman et al., 2014). In a study done by Ranjan et al.

(2018), the algorithm was able to achieve an accuracy of above 94% on three datasets and above 98% on one of the datasets. While AFRS perform well on most of the tasks, they are not accurate all the time. Thus, human oversight is still needed to correct the decisions of AFRS on which they make an error.

Human Face Matching Abilities

While humans are highly accurate at recognising known faces, like that of their friends and family, their accuracy reduces when it comes to matching new faces (Burton et al., 1999). Burton et al. (1999) showed that humans made errors in around 20% of face-matching trials. However, some individuals are better at recognising faces than others. Face matching abilities lie on a spectrum where at the lower end are people who suffer from prosopagnosia (Jones & Tranel, 2001), which is a neurological disorder wherein the individual is unable to detect faces, and on the other end are people called super-recognisers (Russell et al., 2009). Human face-matching abilities can be limited by many reasons, such as low image quality (Bindemann et al., 2013), other race effects (Kelly et al., 2007), and matching faces for long periods of time (Fysh & Bindemann, 2017). These issues can be addressed by the use of AFRS algorithms.

Human and AFRS Interaction

In an ideal scenario where humans and AFRS algorithms are applied together, the operator and the system will agree on the decision when the system is correct and humans will correct the system on the ones that show an error. In other words, the performance of both humans and machines will be optimal However, it has been found that the presence of an AFRS influences the response of the human operator. Howard et al. (2020) demonstrated in their study that showing the participants the identity decisions of an AFRS influences their own judgment in a way that they become more confident that faces labelled ‘same’ are similar and faces labelled ‘different’ are not. This suggests that due to the biased response of

humans, they might not be able to correct the error made by the system, thereby reducing the performance level of the human-automation team.

Additionally, it has, also been shown that the accuracy of human-AFRS team on such tasks is limited due to the individual's performance. In a study by White et al. (2015), participants selected the wrong match from a list of unfamiliar adult faces in 50% of the trials. Another study by Carragher and Hancock (2023), demonstrated that while the performance of humans improved with the assistance of AFRS, the participants were unable to achieve the level of performance that the highly accurate AFRS achieved alone. Such studies suggest that the individual is a limiting factor in this interaction.

Role of Trust in Human-Automation Interaction

Human-automation interaction can sometimes lead to errors, particularly due to the limitation of human operators (White et al., 2015; Carragher & Hancock, 2023). Accidents can occur if the operator *misuses* automation by over-trusting it or *disuses* automation by under-trusting it (Parasuraman & Riley, 1997). For example, in 2012, the Costa Concordia cruise ship sank and killed 32 passengers. Upon investigation, it was discovered that the captain diverged from the ship's computer-programmed route, due to under-trusting the ship's navigation system, before it hit the shallow reef that caused the sinking (Levs, 2012). In 2009, Turkish Airlines Flight 1951 crashed killing 9 people. It was partially caused because the pilots continued to rely on the autopilot system, despite the failure of an altitude-measuring instrument (CNN, 2009). These accidents are two of many that may have occurred due to *misuse* and *disuse* of automation. Thus, it becomes essential to instil appropriate levels of trust in automation in operators to increase the safety and productivity of human automation teams.

Furthermore, trust and reliability in automation have been found to be related to each other (Rice, 2009). Meyer (2004) shows that the level of trust in automation is positively

correlated with reliance on automation, which means low levels of trust will correspond to low levels of reliance. Another study by Lee and Moray (1994) suggests that operators are more likely to use the automated system when their trust in the system exceeds their self-confidence in the task. Increased reliance on the decisions made by automation, when the system performs better than the operator, will lead to improved interaction between humans and automation and reduce the number of errors.

Factors that Influence Trust in Automation

Hoff and Bashir (2015) describe that the variability in human-automation trust is due to the human operator, the environment, and the automated system. This reflects the different layers of trust respectively as identified by Marsh and Dibben (2003), that is, '*dispositional trust', 'situational trust', and 'learned trust'*. The individual's tendency to trust the automation is the '*dispositional trust'*, while the environment of the user majorly influences the '*situational trust'*. '*Learned trust'* depends on the operator's past experiences with the automated system. Hoff and Bashir (2015) described that under situational trust, there are two broad sources of variability: the external environment and the internal environment, which are the context-dependent characteristics of the human user. The type of system, system complexity, task difficulty, perceived risks, and perceived benefits are some of the external variability factors that can influence an operator's trust in the automated system (e.g., Bailey & Scerbo, 2007; Fan et al., 2008; Madhavan er al., 2006; Ross, 2008; Spain, 2009). Perceived task difficulty also influences the human's trust and reliability in automation (Parkes, 2009; Schwark et al., 2010).

Perceived Task Difficulty as a Factor Influencing Trust in Automation

As described in the study by Hoff and Bashir (2015), one of the external variability factors that contributed to trust in automation is perceived task difficulty. According to Madhavan et al. (2006), human operators evaluate the capabilities of automated systems

based on the relative difficulty of the task. Madhavan et al. (2006) aimed to test whether failures by automated aids on tasks that were easily performed by a human operator would undermine trust in automation, by conducting a signal detection task. The participants were divided into three groups, two groups were provided with decision aids and the third group was unaided. However, the aid for half of the participants only made errors on easy targets while for the other half, it only missed on difficult targets. As hypothesised, participants who utilized the aid that only missed easy targets had lower trust in automation than participants who utilised the aid that only missed the difficult targets. Participants in the easy miss group also disagreed with the aid on approximately 50% of the difficult target trials, even though the aid was perfectly accurate on these trials. This suggested that the resulting lower levels of trust also led the participants to overrule the correct decisions of the decision aid, thus increasing the likelihood of errors.

Schwark et al. (2010) conducted an experiment to find out how perceived task difficulty influenced reliance and compliance with automation. The participants were asked to detect the letter 'X' in a number of random letters and were informed that the computer aid that was provided for assistance was not perfectly accurate and the exact accuracy was unknown. Before each trial, participants were told how hard the trial would be. The results of the study suggested the participants were more likely to utilise the aid provided to them in trials that they perceived as hard, indicating that perceived task difficulty influences the way a human operator relies on automation.

It has also been suggested that humans are incapable of making optimal decisions in a situation due to the amount of effort required to do so (Simon, 1995). As a result, when humans make decisions in tasks of high difficulty, they are more willing to reduce effort while maintaining accuracy in an attempt to reduce the workload to a manageable level (Payne et al., 1993). Thus, when an operator perceives the task as easy, they would be

unlikely to utilize decision aids (Parkes, 2013). Thus, Parkes (2013) hypothesized that the use of decision aid is positively related to the perceived difficulty of the task. In the experiment, the participants read through the case study about a company and recorded two decisions regarding the company's future, one unaided and one aided. They then answered a few questions related to the case and the decision aid. Both the reliance on decision aid and perceived task difficulty were self-assessed by the participants. The results showed that perceived task difficulty is a significant positive predictor of reliance and the use of decision aid.

Moreover, it has been found that during decision-making, people generally overestimate their competence in tasks, especially if the task appears easy to them (Dunning et al., 2003). When applied to human-automation interaction, it suggests that people will tend to rely on their own decisions, as opposed to the decision of the automated aid, if the task seems easy to them as compared to when it appears difficult. All these studies suggest that perceived task difficulty influences an operator's trust and reliance on automation.

Current Study

The current study seeks to extend the research done on the perceived task difficulty and reliance on automation to reliance on Automated Facial Recognition Systems (AFRS) specifically. As aforementioned, facial recognition technology is used by passport officers, border control officers, and security and defence personnel. By understanding the factors that influence their reliance on the AFRS, the efficiency of their interaction and the decision quality will increase.

The current research aims to find how the perceived level of difficulty affects trust and reliance on AFRS in face-matching tasks. Based on these previous studies, we hypothesized that:

1. Participants who are assigned to the ‘hard’ perceived difficulty condition will change the degree of confidence in their response to match with the AFRS more often compared to those in the ‘easy’ perceived difficulty condition, thus showing a higher degree of reliance.
2. The accuracy of participants who were informed that the task would be “hard” will improve more with the assistance of AFRS as compared to the participants in the “easy” condition.
3. Participants who respond that they found the task ‘hard’ in the post-task questionnaire, regardless of the task condition they were randomly assigned to, will also change their initial response to match with the AFRS more as compared to participants who find it ‘easy’.
4. The relationship between perceived task difficulty and change in the degree of confidence in their response to match with the AFRS will be stronger when trust in the AFRS is higher.

Method

Sample Size

To determine the number of participants in the study, we performed an a priori analysis using G-power (Erdfelder, Faul, Buchner & Lang, 2009). With an anticipated moderate effect (Cohen's $d = 0.5$) and the desired power of 80%, the recommended sample size was 128 (total), with 64 in each condition. For the purpose of our study, the moderate effect is the smallest effect of interest. Since the study has exclusion criteria, we aimed for a total sample size of 128 participants, with 64 in each group, after the exclusion of participants.

Participants

Participants ($n = 53$) were first-year psychology students enrolled at the University of Adelaide, who participated in the study in exchange for course credit. Our study had a few exclusion criteria. Participants who completed the face-matching task too quickly (in less than seven minutes), completed the task too slowly (more than 60 minutes), did not complete the task, or completed it more than once were excluded from the analysis. Additionally, participants who answered incorrectly on the manipulation check question which stated the perceived level of difficulty shown to them, failed either of the two attention check face matching trials or answered incorrectly on the attention check question at the end of the task about the stated accuracy of the AFRS were also excluded. 5 participants failed the manipulation check, 1 participant took too long to complete the task, 3 participants started the task but did not complete it and 1 participant did the task twice. Of the 43 remaining participants, 36 were females and 7 were males. The ages of the participants ranged from 18 years to 43 years ($M = 19.25$ years; $SD = 3.68$).

The study was approved by the Human Research Ethics Committee of the University of Adelaide (HREC-2019-23/01). Informed consent was obtained from the participants before the study.

Design

Our study is a mixed measures experiment. There is a between-subjects factor of perceived task difficulty (easy and hard), which is manipulated at the start of the experiment. The within-subjects factor in the study is the use of the AFRS (unassisted decision and assisted decision).

Materials

Face Matching Abilities

Before starting the face-matching task, the participants were shown 4 questions that were aimed at understanding their initial confidence in their own accuracy on the task. An example of the question is '*How confident are you in your ability to accurately judge whether two photographs show the same person*?' (Extremely High to Extremely Low). This gave us an understanding of how well they think they are going to perform on the task and how likely they are to use the assistance of the AFRS before the manipulation.

Perceived Difficulty Manipulation

As the study aimed to examine whether perceived task difficulty would affect the reliance and trust of participants on the AFRS, we manipulated how the participants would perceive the level of difficulty of the task. At the beginning of the experiment, participants were randomly assigned to one of two perceived difficulty conditions: 'Easy' or 'Hard'. Each condition was shown a unique statement before the face-matching task. The statement was designed in a way that would lead the participants to believe that the task is either easy or difficult. These statements are as follows:

1. ‘Easy’ condition: *Face matching is actually a surprisingly easy task! Many people find it very easy to achieve accuracy above 90% correct, even without any external assistance.*
2. ‘Hard’ condition: *Face matching is actually a surprisingly difficult task! Many people find it very hard to achieve accuracy above 60% correct, without any external assistance.*

People vary on the levels of their face-matching abilities. Some people do well, achieving around 90% accuracy on the Glasgow Face Matching Test 2-S, while others perform slightly lower, achieving around 60% accuracy on the Glasgow Face Matching Test 2-S (Burton et al., 2010) Thus, to present the true performance of people on the task, we chose to show these accuracy scores to the participants for the perceived difficulty manipulation.

Text-Based Manipulation Checks

To check whether the participants read the instructions for the manipulation correctly, the manipulation statement was followed by the question, “*In a face matching task, many people find it very…”,* with the possible answers as “*hard to achieve accuracy above 60% correct”* or *“easy to achieve accuracy above 90% correct”.* Participants who answered incorrectly on the manipulation check were excluded from the study.

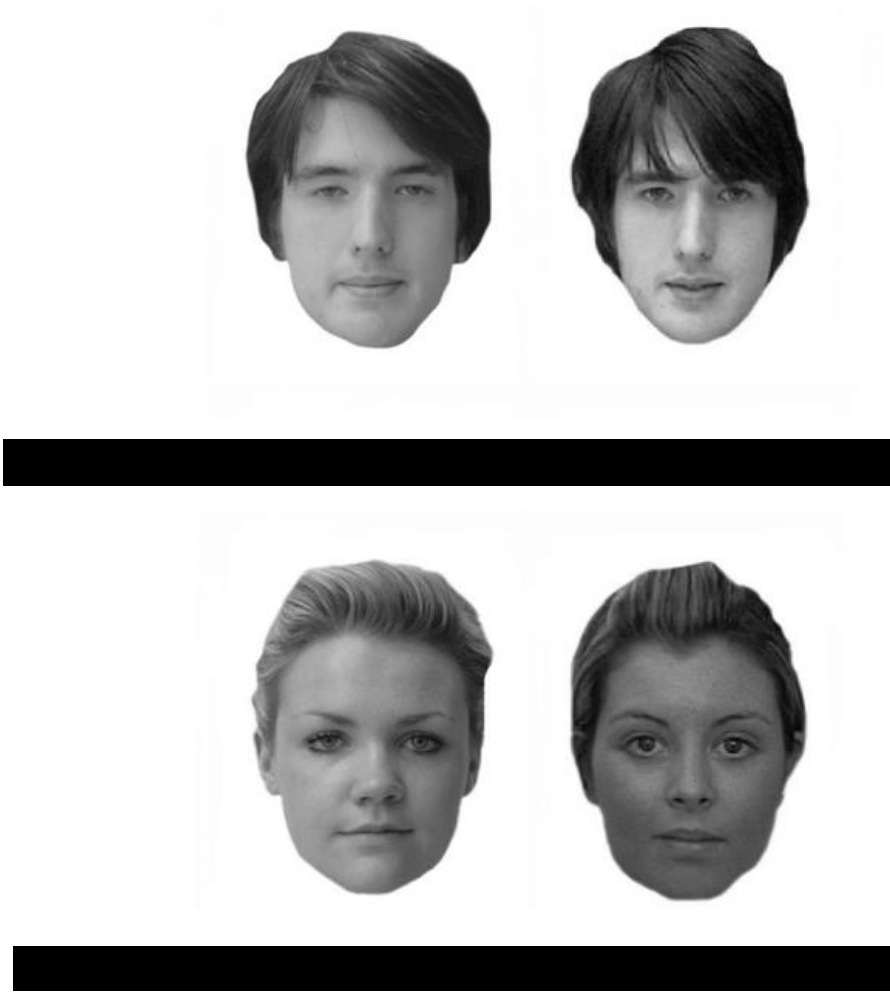
Glasgow Face Matching Test-2 Short Version (GFMT2-S)

The short version of the Glasgow Face Matching Test 2 (GFMT2-S; White et al., 2021) was used to measure an individual’s face-matching ability. The test items show pairs of portrait images of people’s faces that either match (same person) or don’t match (different people). The faces are shown simultaneously next to one another, and the participant is asked whether the faces are the same person or two different people. The images vary with respect to the head angle, pose, expression, and subject-to-camera distance. The GFMT2-S has 80 items and comprises of two sets of 40 items each of equal difficulty. Figure 1 shows an

illustration of what the image pair types look like in GFMT2-S in our study. Internal consistency of the test was favourable (Cronbach's alpha = .94). The test also shows high test re-test reliability (r = .78) (White et al., 2021).

Figure 1

Examples of the Image pairs used in the GFMT2 in our study



Note. These images are to show representative test items that do not contain the people appearing in the test (White et al., 2021).

Attention Check Face Matching Trials

To check whether the participants were paying attention during the task, they were shown images of the faces of two different celebrities at two different points in time during the task.

If they answered incorrectly on either of the attention check face matching trials, they were excluded from the study.

Automated Facial Recognition System

The decisions made by the simulated AFRS in our study were based on the performance of a real AFRS on the GFMT2-S (see Carragher & Hancock, 2020). The AFRS made a final decision by computing a similarity score between the two faces and evaluating it against a threshold value (see Carragher & Hancock, 2022, 2020). The simulated AFRS in our study was accurate in 95% of the trials.

Trust Questionnaires

The questionnaires used in the study were a combination of the published Lee and Moray's (1994) questionnaire and questions adapted from Carragher, Sturman and Hancock's (In Prep) study.

Lee and Moray Questionnaire. Lee and Moray's (1994) measure was used to evaluate the trust of the participants in the AFRS. The measure has two items: one is concerning the trust in automation, whereas the other measure is regarding the participant's self-confidence. The ratings are made on a scale from 0 to 10 ("extremely low" to "extremely high").

Exploratory Questionnaire. Exploratory questions about the participant's trust in the system were also asked as a part of the experiment. These questions were adapted from Carragher, Sturman, and Hancock's (In Prep) study. The questions were related to the participant's trust in AFRS and were asked both pre and post-experiment. Some questions were also aimed at understanding the participant's perception of their own ability in relation to that of the AFRS The possible responses to the questions ranged from being in the form of "yes or no" to being on an 11-point Likert scale, multiple-choice, or being measured as a percentage. Some example questions and their possible responses were "*Do you trust the facial recognition system to help you in this task?*" (Yes or No), "*How challenging do you*

think this task is going to be?" (11-point Likert scale, from "extremely difficult" to "extremely easy"), "*who do you think will be more accurate at this task*?" ("You" or "the Facial Recognition System"), and "*how accurate do you think you are when completing the task on your own (unassisted)*"? (0-100%).

Procedure

The study began with obtaining the participants' informed consent and basic demographic information like age and gender. The participants were then given general instructions about the face-matching task and the 6 possible responses that they could give, '*Definitely Same', "Probably Same', "Guess Same',* '*Guess Different', 'Probably Different',* and *'Definitely Different'.* They were then asked four questions to examine their initial understanding of their ability on the task.

The participants were then randomly assigned to the two task difficulty conditions: '*Easy'* and *'Hard'.* The perceived task difficulty was manipulated by showing two distinct statements as mentioned above, that were designed to make the participants believe that the task was either going to be easy or hard. Following the manipulation, the participants went through a manipulation check wherein they were asked what percentage of accuracy (60% or 90%) they were shown on the previous screen. Participants who responded incorrectly to the question were excluded from the study. They were then shown detailed instructions of the task with example trial displays to show the participants how AFRS decisions look in action.

Before they started with the main task, they were asked 16 pre-task questions that helped understand how the participants thought they might use the simulated facial recognition systems while making a decision (see Appendix 1).

Participants then started the face-matching task. First, participants judged whether the pair of images shown in each trial are the same or different, without the assistance of the AFRS (Unassisted decision). After they made their initial response, they were shown the

decision of the simulated AFRS. While the participants could give 6 possible responses, the AFRS only showed its response as either ‘Same’ or ‘Different’. The participants were then asked for their final decision, that is with the assistance of the AFRS, on whether they thought that the pair of face images were the same person or different people (Assisted decision).

After they finished the main task, they were again asked 17 post-task questions about how they tried to use the simulated facial recognition system into their decisions. This concluded the task for the study.

Analysis

Manipulation Check

In order to assess whether the manipulation worked, the participants were asked “*How challenging do you think this task is going to be?*’ before they were assigned to the task conditions and after they were assigned to the task conditions. The scores were on a 0-10 scale, where 0 meant easy and 10 meant hard.

Accuracy

The overall accuracy on the task was measured by dividing the ‘number of correct trials’ by the total number of trials (n = 80) multiplied by 100 to obtain a percentage. This was used for both initial and final accuracy. For the convenience of measurement, participants were considered to be accurate if they responded on the correct trials of 'Same’ faces as any of the three responses of ‘Definitely Same’, ‘Guess Same’, and ‘Probably Same’, and similarly for the correct trials of the ‘Different’ faces.

Change in the Degree of Confidence

Reliance on the AFRS was measured by the change in the degree of confidence of the participant’s final response to match with the AFRS. The 6 responses were based on a scale, with ‘Definitely Different’ being 1 and ‘Definitely Same’ being 6. The identification decision

made by the AFRS only specified if it was ‘Same’ or ‘Different’, which we classified as ‘Definitely Same’ and ‘Definitely Different’ for the convenience of measurement. Any change of the participant’s response from the initial identification decision to the final identification decision was measured in units. If they changed their decision to match with the AFRS, it was measured in positive values and showed *reliance* on the system, whereas if they changed their decision against the AFRS, it was measured in negative values, which showed *rejection* of the system. For example, if a participant’s initial response was ‘Guess Same’(measured as 4 on the scale), and the decision made by the AFRS was ‘Same’, and they changed their final response to ‘Probably Same’ (measured as 5) and showed reliance on the system, they moved +1 unit. However, if they moved from their initial response to ‘Guess Different’ (measured as 3) which goes against the decision of the AFRS and shows rejection, it was measured as -1 unit.

Trust in the AFRS

Trust in the AFRS was measured using Lee and Moray’s (1994) trust/self-confidence measure. The measure contains two items, self-confidence ratings and trust in the system rating. The self-confidence ratings were subtracted from trust in the system ratings. This resulted in a relative change score, in which positive values indicated higher trust in the AFRS and negative values meant higher self-confidence.

Statistical Calculations

All the statistical tests were run using SPSS-27. The same software was used to graph the data.

Results

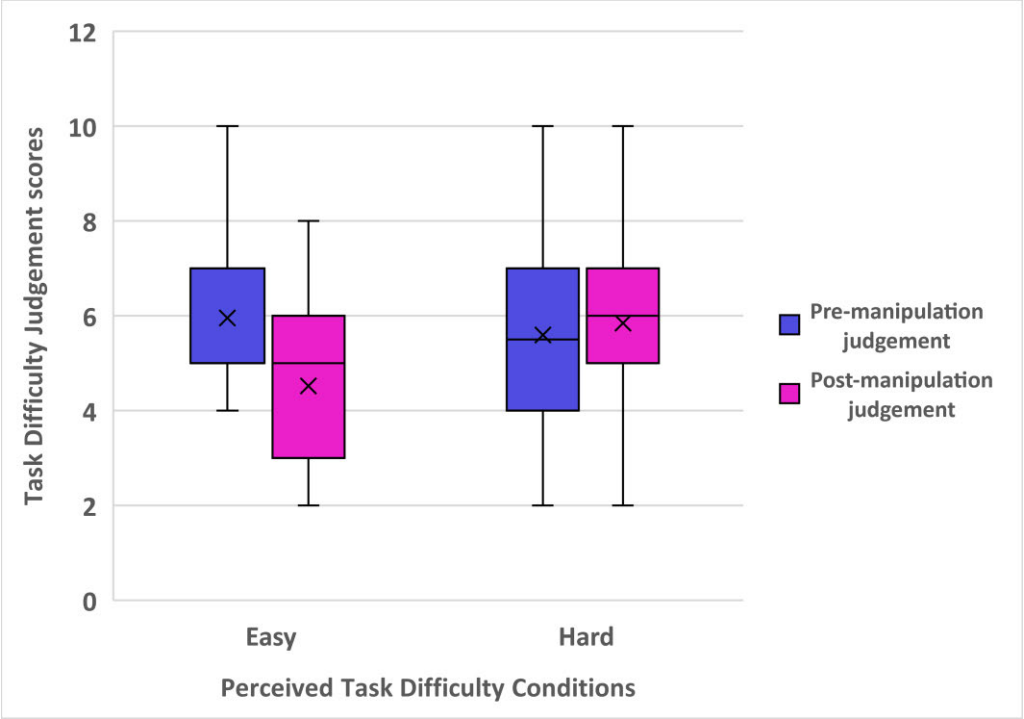
Manipulation Check

We performed a two-way mixed ANOVA with pre-manipulation judgement of difficulty and post-manipulation judgement of difficulty as within-participants factor and perceived task difficulty condition as the between-participants factor (easy and hard). The test was done to check whether there was a difference in the means of difficulty judgement pre and post manipulation between the two perceived task difficulty conditions, which would then suggest that the manipulation worked in the way we expected it to. The mean of task difficulty scores in the 'Easy' condition was 5.96 (SD=1.82) before manipulation, which lowered down to 4.52 (SD=1.59) after manipulation. The mean of task difficulty scores before manipulation in the 'Hard' condition was 5.6 (SD=2.14), which increased slightly to 5.85 (SD=1.98).

There were no outliers in our data, as assessed by the visual inspection of the boxplot (see Figure 2). Shapiro-Wilk test suggested non-normal data in the pre-manipulation judgement in '*Easy'* condition (p = .007). However, we decided to use the same method as ANOVAs are considered to be robust to deviations from non-normality (Sawilowsky & Blair, 1992) The assumption of homogeneity of variances and homogeneity of covariances was met, as per Levene's Test of Homogeneity of Variances and Box's M test respectively (p > .05).

Figure 2

Boxplot of judgment scores pre-manipulation and post-manipulation in two task conditions



Note. The cross in the boxes denotes the mean of the dataset. The black line in the box marks the median of the dataset. The lower whisker shows the minimum value, whereas the upper whisker shows the maximum value.

There was a statistically significant interaction between the perceived task difficulty and manipulation on judgement scores of difficulty, $F(1,41) = 7.81$, $p = .008$, partial $\eta^2 = .16$. This suggests that the participants believed the task was going to be easier after manipulation than previously thought and the participants in the hard condition believed that it would be harder after the manipulation, with a large effect size. The main effect of the perceived task difficulty conditions (easy and hard) was statistically significant in task difficulty judgements scores, $F(1,41) = 5.95$, $p = .019$, partial $\eta^2 = .127$. The main effect of manipulation (before and after) was statistically significant in task difficulty judgement scores in the easy condition, $F(1,22) = 11.37$, $p = .003$, partial $\eta^2 = .341$. However, the main effect of

manipulation (before and after) was not statistically significant in task difficulty judgement scores in the hard condition, $F(1,19) = 3.51$, $p = .561$, partial $\eta^2 = .018$.

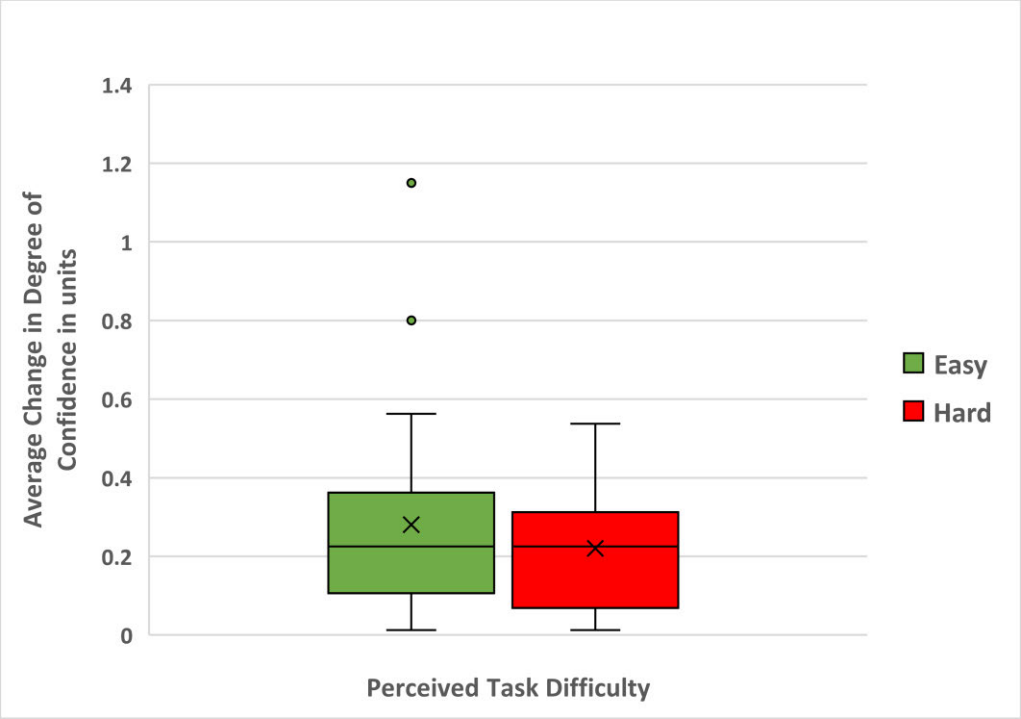
Hypothesis 1: Change in Degree of Confidence to Match with the AFRS in the Two Perceived Difficulty Task Conditions

Change in degree of confidence of each condition was examined at a group level. The change in degree of confidence of the participants in easy condition was very similar to that of the hard condition (Easy: M=0.27, SD= 0.26; Hard: M=0.26, SD= 0.16). Since the mean values are positive, it indicates that both the groups relied on the decision of the AFRS more often than rejecting the decision. We checked for normality of the data using a Shapiro-Wilk test. Significant results in the Shapiro-Wilk test suggest that there is evidence of non-normality in the easy condition ($W = 0.77$, $p < .001$).

We also checked for any outliers in the data. On visual inspection of the boxplot, we can see that there were two outliers in the ‘easy’ group and one in the ‘hard’ group (see Figure 3). Since the sample size we had at the time of analysis was small, it is hard to say that the outliers in our study are true outliers. Due to this reason, we did not exclude the outliers in our sample and included them as normal. Moreover, the direction of the outliers in our sample are minimising the differences in the two task conditions.

Figure 3

Average change in degree of confidence in units in two task conditions



Note. The dot symbols in the graph show the outliers in our sample. The cross in the boxes denotes the mean of the dataset. The black line in the box marks the median of the dataset. The lower whisker shows the minimum value, whereas the upper whisker shows the maximum value.

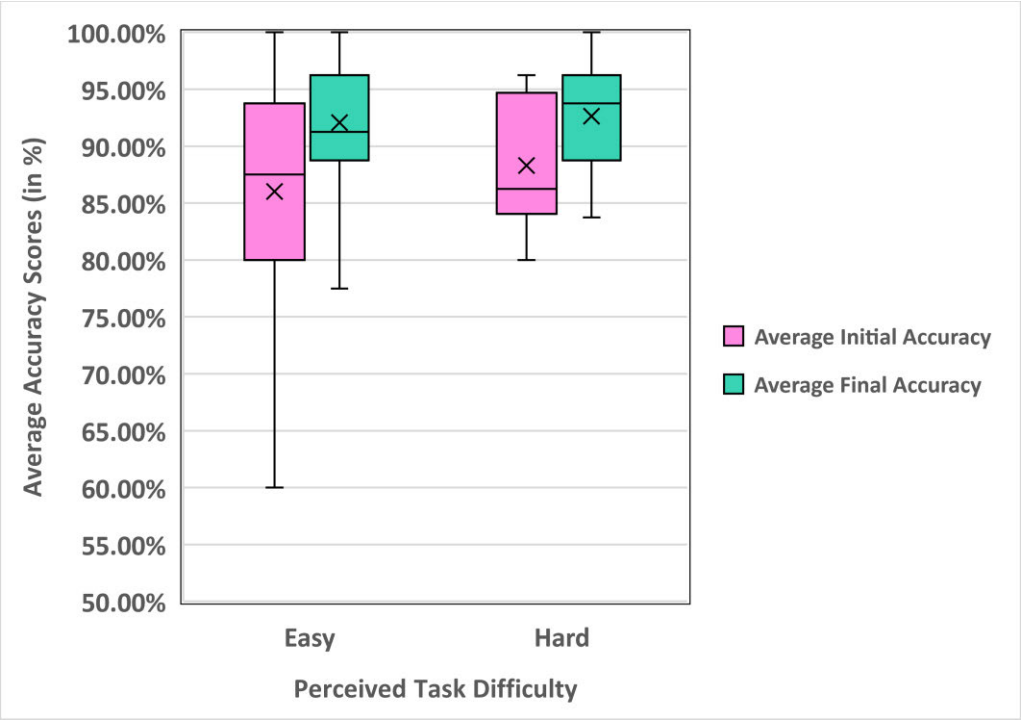
As the data failed the assumptions of an independent samples *t*-test, we decided to use a non-parametric Mann-Whitney Test to analyse the data for the first hypothesis. The results indicated that there was no significant difference between the reliance shown by participants in 'easy' condition (*Mdn* = 0.20) and reliance shown by participants in 'hard' condition (*Mdn* = 0.26), *U*=246, *Z* = 0.39, *p* = .697. The effect size was calculated using the rank biserial correlation, which indicated an effect size of -0.07. This suggests a small effect size for the data.

Hypothesis 2: Change in Initial and Final Accuracy in Participants in the Two Perceived Task Difficulty Conditions

Participants' accuracy scores were submitted to a 2 X 2 repeated measures ANOVA with initial (M= 87.09, SD= 7.65)and final accuracy scores (M= 92.32, SD= 5.49)as a within-participants factor and the between-subjects factor of perceived task difficulty (easy or hard). The test was performed to investigate whether the accuracy of participants in the 'hard' condition improved more with the assistance of the AFRS as compared to the 'easy' condition.

Figure 4

Boxplot of the average accuracy scores in the 'Easy' and 'Hard' task conditions.



Note. The cross in the boxes denotes the mean of the dataset. The black line in the box marks the median of the dataset. The lower whisker shows the minimum value, whereas the upper whisker shows the maximum value.

The assumption of homogeneity of variances in the average initial accuracy was not met as per Levene's Test of homogeneity of variances (p = .049). The data also did not meet

the assumption of covariances as per Box's M test ($p = 0.39$). On inspection of the boxplot (see Figure 4), there were no outliers in our sample.

There was no statistically significant difference in the interaction term. There was no significant difference in the improvement of the accuracy scores between the participants in two conditions, $F(1,41) = 1.13$, $p = .293$, partial $\eta^2 = .004$. The main effect of assistance of the AFRS showed a statistically significant difference in accuracy scores at initial and final level, $F(1,41) = 41.03$, $p < .001$, partial $\eta^2 = .50$. This means that the participants improved on their accuracy after the assistance of the AFRS. The main effect of Perceived task difficulty showed that there were no statistically significant differences in accuracy scores between the two task conditions, $F(1,41) = .57$, $p = .454$, partial $\eta^2 = .014$.

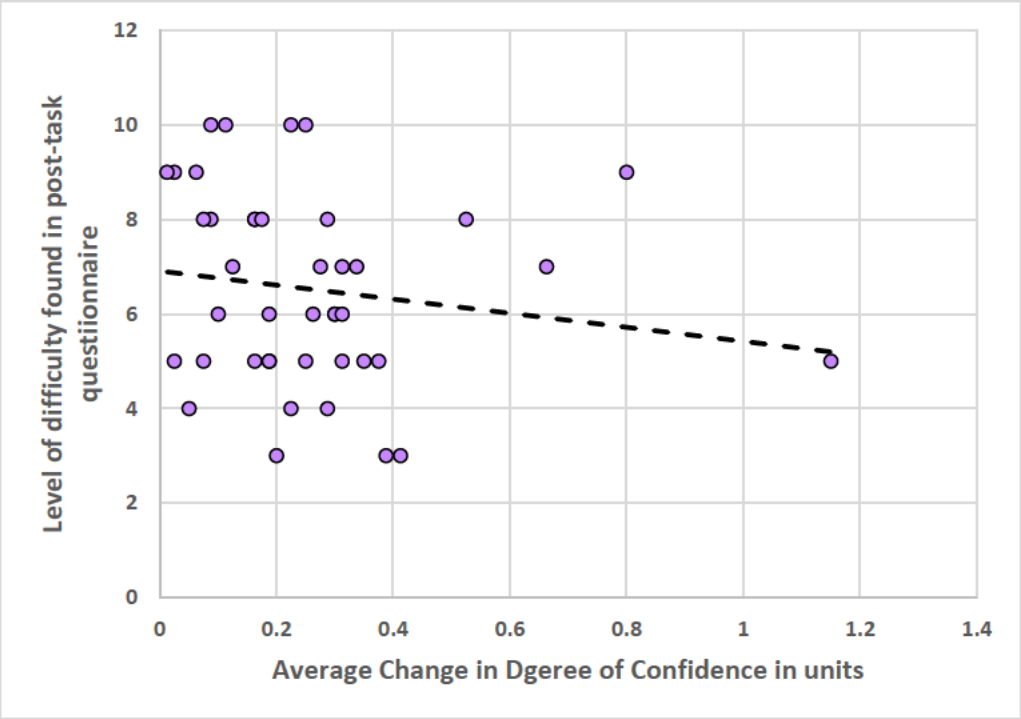
Hypothesis 3: Relationship Between Reliance and Self-assessment of Task Difficulty on Post-Task Questionnaire

A correlation test was performed to examine the relationship between reliance on the AFRS and self-assessment of task difficulty as reported by the participants in the post-task questionnaire. The response was measured on a 0-11 likert scale, with 1 being '*extremely difficulty'* and 11 being '*extremely easy'.* Preliminary analyses showed that relationship was linear, however, there was evidence of non-normality in the data ($W=0.81$, $p<.001$). We decided to proceed with the Pearson's correlation test as it has been found to be somewhat robust to non-normality (Kang & Harring, 2012).

The relationship between the reliance of participants and how hard they found the task to be in the post-task questionnaire was not significant, $r(41) = -.16$, $p = .321$, with the task difficulty rating in the post-task questionnaire explaining 10.3% of the variability in reliance shown by the participants. The Pearson's *r* depicted a negative small effect, which means as reliance increases, participants answered the task to be easier in the post-task questionnaire (see Figure 5).

Figure 5

Scatter plot of reliance on the AFRS and self-assessement of task difficulty in the post-task questionnaire



Note. The points on the graph show the subjective task difficulty of the participants corresponding to their average change in degree of confidence. The dotted line shows the trendline of the data.

Hypothesis 4: Effect of Trust on the System on Perceived Task Difficulty and Reliance

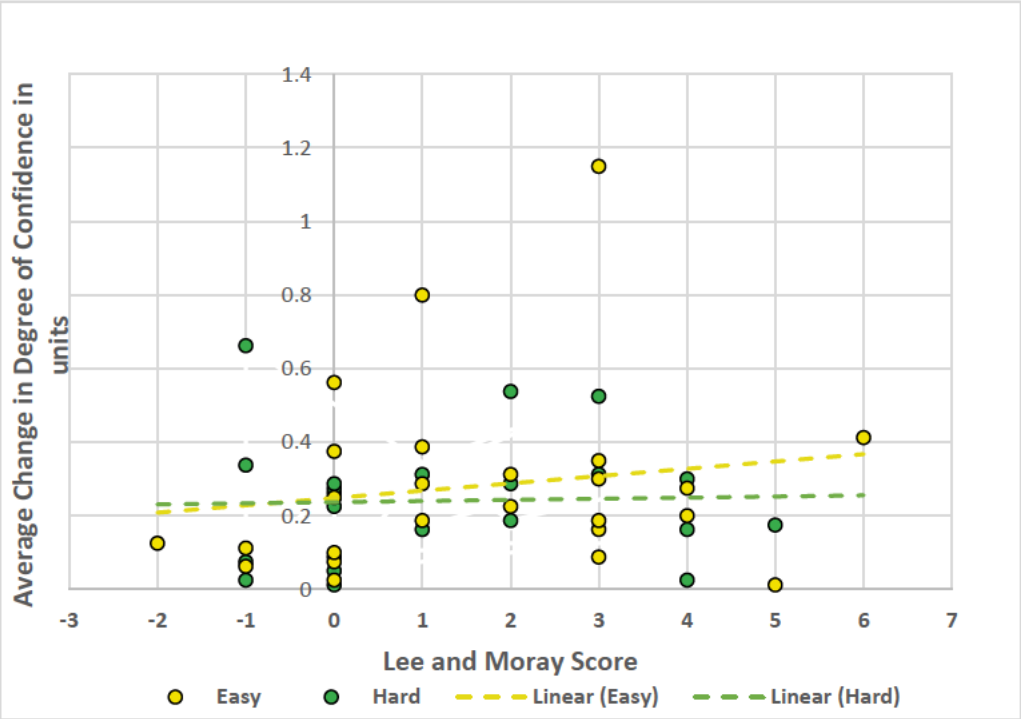
A one-way ANCOVA was conducted to determine the effect of perceived task difficulty on change in degree of confidence in units, after controlling for the participants' trust in the system. The data showed evidence of non-normality. However, ANCOVA has been found to be robust to non-normality (Osborne & Overbay, 2008). There was homogeneity of variances, as assessed by Levene's test of homogeneity of variance (p=.407). On visual inspection of the scatterplot, there was a linear relationship between trust in the system and reliance. The slopes for both the task conditions were positive.

After adjusting for the trust of the participants in the system, no statistically significant interaction was found between the reliance in the system and perceived task difficulty conditions, $F(1,40)=0.002$, $p = .962$, partial $\eta^2=.0.00$. The main effect of trust in system on reliance on the system was not statistically significant, $F(1,40) =.45$, $p = 0.506$, partial $\eta^2=.0.11$.

Figure 6 shows the relationship between trust in the system using Lee and Moray score and average change in degree of confidence in the participants in the 'easy' condition and the 'hard' condition.

Figure 6

Relationship between trust in the system and on average change of degree of confidence in the responses of participants in the 'easy' and 'hard' condition.



Note. The points on the graph show the Lee and Moray trust scores of the participants corresponding to their average change in degree of confidence. The yellow dotted line shows the trendline of the 'Easy' condition. The green dotted line shows the trendline of the 'Hard' condition.

Discussion

The aim of this study was to examine whether perceived task difficulty of the face matching task had any effect on the participants' reliance on the AFRS. While we did not find any significant differences between the reliance of the two perceived task difficulty group, we did find that in general participants relied on the system more, that is moved towards the decision of the AFRS, instead of rejecting it. Howard et al. (2020) showed similar findings in their study that participants become more confident that the faces labelled 'same' are similar, and faces labelled 'different' are not. Our study also supported previous studies like the one by Carragher and Hancock (2023) that the use of AFRS as a decision aid does improve human performance in face matching tasks. However, the accuracy of human-AFRS team does not exceed to the accuracy of the AFRS alone. This provides further evidence that humans could be a limiting factor in the interaction (White et al., 2015; Carragher & Hancock, 2023).

In our study, we manipulated the perceived task difficulty by randomly allocating participants into two conditions, '*Easy'* and *'Hard',* and showed them two distinct statements that suggested if the task they were going to do is easy to score on or difficult to score. In order to check whether our statements truly changed the perception of task difficulty as we intended it to, we performed a manipulation check. Significant results suggested that the manipulation made the participants in the easy condition believe that the task was going to be easy, and participants in the hard condition believed that the task was going to be difficult. This shows us that the manipulation could be an attributing factor to our results.

The first hypothesis looked into whether participants in the 'Hard' condition would rely more, or change the degree of confidence of their response in order to match with that of the AFRS more often, as compared to the participants in the 'Easy' condition. Contrary to our expectations, reliance on AFRS did not differ between the 'Easy' and 'Hard' conditions. This finding was contrary to the findings of Hoff and Bashir (2015) and Schwark et al (2010) who

suggested that perceived task difficulty influences trust and participants were more likely to utilize a decision aid when they found the task to be hard. A possible explanation for this result could be that as participants performed on more and more trials, they made their own perceptions of task difficulty, which might have not matched with the task difficulty condition that they were assigned to. Parkes (2013) showed in their study that participants relied more on the decision when they found the task to be difficult in their own perception. Additionally, the operator's decision to rely on the decision aid depends on their trust in the system and self-confidence in their own ability (Lee & Moray, 1994). When trust in the system is greater than their self-confidence, greater reliance in automation can be observed. However, sccording to Dunning et al. (2003), people are more likely to overestimate their abilities. As a result, participants who overestimated their skills and had greater self-confidence in the face matching task might Another explanation could be that perceived task difficulty is not the same as the actual difficulty of the task. Since all the participants in the experiment performed the same test, suggesting same task difficulty across the two perceived task difficulty conditions, it could have influenced our results.

Hypothesis 2 stated that participants in the 'Hard' condition would show more improvement in their initial and final accuracy scores, as compared to the 'Easy' condition. This was hypothesized on the basis that when participants would rely more on the decisions of the AFRS, their accuracy would increase as a result. Our results did not support this hypothesis and showed that there was no difference in the improvement in accuracy scores across the two perceived task difficulty conditions. This could be due to the non-significant results found in Hypothesis 1. Since there was no difference between the reliance of the two difficulty conditions, it could have led to no difference in the accuracy scores of the two difficulty conditions. The results, however, did indicate that the accuracy scores of the participants across the two task conditions improved with the assistance of the AFRS. This finding is consistent

with the findings of Carragher and Hancock (2023), who also found AFRS to be a contributing factor in the improvement of human performance on face matching tasks.

According to the third hypothesis, participants who found the task 'Hard' in the post-task questionnaire, regardless of the perceived task difficulty condition they were randomly assigned to, would also rely on the AFRS more as compared to those who found the task to be 'Easy'. The results did not support this. No significant relationship was found and there was a suggestion of a small negative effect. This suggested that people who self-assessed the task to be 'easy' in the post-task questionnaire, relied on the AFRS more. Our finding contradicts the study by Parkes (2013), which indicated that perceived task difficulty, when self-assessed by the participants, positively correlated to the use of decision aids. One possible explanation is that the participants reliance in the AFRS could have attributed to the participants' self-assessment of the task as 'easy'. This means as they actually found the task to be easy by the end of it because of their continued use of the AFRS.

The fourth hypothesis assumed that there would be a stronger relationship between perceived task difficulty and reliance on the AFRS when trust in the AFRS was stronger. The results of our study did not support this hypothesis, which was in contrary to the findings of Meyer (2004). According to Meyer (2004), levels of trust in automation is positively correlated to reliance in automation. This raises questions about the relationship between the participant's trust in automation and in turn, reliance in automation.

There were a few limitations in our study. One of them is that at the time of analysis, the sample size we received was very small. At the beginning of our study, we conducted an a priori analysis, using G-power (Erdfelder, Faul, Buchner & Lang, 2009), to determine the sample size. It was indicated that we needed a sample size of 128 participants, with 64 participants in each condition, however, we only managed to get a total of 43 participants, 23 in the 'Easy' condition and 20 in the 'Hard' condition. While this is an ongoing study, the small

sample size could have had various implications on the interpretation of our results. It becomes difficult to detect any true differences between groups when the sample size is small (Akobeng, 2016). Human face matching abilities lie on a spectrum, where on one end are people who are unable to detect faces (Jones & Tranel, 2001) and on the other end, there are people who perform exceptionally well at such tasks (Russell, Duchaine, & Nakayama, 2009). In a small sample size, it becomes difficult to gather estimates of true performance of the population. As a result, the statistical power of our study was reduced due to a smaller number of participants. Further research is needed to examine whether our findings are true and valid, and that a Type II error was not made.

Our study has various implications in applied settings. Since the AFRS is being used in various work environments, such as border control, where passport officers use e-gates for verification purposes (Noyes & Hill, 2021), our study looks into one of the many factors that could influence the use of AFRS to make a decision. Our results are based on one experiment and further investigation is required to validate them and explore whether perceived task difficulty would affect reliance on the AFRS as it has been shown to do with other decision aids (Schwark et al., 2010), Parkes (2012), Hoff & Bashir (2015). Further research is also needed to find whether relying of the decision of the AFRS would increase human accuracy in face matching and reach an optimal level of performance. This would help form and implement strategies that would help workplaces that use AFRS, such as border control.

In conclusion, our study did not find that perceived task difficulty influences reliance of human operators in AFRS. However, it does support the findings that humans are more likely to rely on the AFRS rather than rejecting it. The tendency to rely on the AFRS, in turn, increases human accuracy on face matching tasks, but the performance is on a sub-optimal level. These results strongly suggest that there are other factors at play, possibly pre-existing biases, and attitudes, that have a stronger influence on human reliance in automation. Further

investigation in this field can help determine how we can reach an optimal level of performance by influencing human operator's reliance in automation.

References

Akobeng. (2016). Understanding type I and type II errors, statistical power and sample size. ACTA PAEDIATRICA, 105(6), 605–609. h

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. Theoretical Issues in Ergonomics Science, 8(4), 321-348.

Bindemann, Attard, J., Leach, A., & Johnston, R. A. (2013). The Effect of Image Pixelation on Unfamiliar-Face Matching. Applied Cognitive Psychology, 27(6), 707–717.

Burton, White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. Behavior Research Methods, 42(1), 286–291.

Burton, Wilson, S., Cowan, M., & Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence from Security Surveillance. Psychological Science, 10(3), 243–248.

Carragher, & Hancock, P. J. B. (2020). Surgical face masks impair human face matching performance for familiar and unfamiliar faces. Cognitive Research: Principles and Implications, 5(1), 59–59.

Carragher, & Hancock, P. J. B. (2023). Simulated Automated Facial Recognition Systems as Decision-Aids in Forensic Face Matching Tasks. Journal of Experimental Psychology. General, 152(5), 1286–1304.

Carragher, Sturman, & Hancock (In Prep). Trust in Simulated Automated Facial Recognition Systems.

CNN Europe (2009, March 4). *Faulty Reading Helped Cause Dutch Plane Crash.* CNN. http://edition.cnn.com/2009/WORLD/europe/03/04/plane.crash/index.html#:~:text=Because%20of%20the%20%22faulty%22%20left,it%20was%20at%20ground%20level .

Dunning, Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why People Fail to Recognize Their Own Incompetence. Current Directions in Psychological Science : a Journal of the American Psychological Society, 12(3), 83–87.

Fan, Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. (2008). The influence of agent reliability on trust in human-agent collaboration. ACM International Conference Proceeding Series, 369, 1–8.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149-1160.

Fysh, & Bindemann, M. (2017). Effects of time pressure and time passage on face-matching accuracy. Royal Society Open Science, 4(6), 170249–170249.

Grother, P., Ngan, M., & Hanaoka, K. (2019). Ongoing face recognition vendor test (FRVT) part 1: Verification. *National Institute of Standards and Technology*.

Hoff, & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors, 57(3), 407–434.

Howard, Rabbitt, L. R., & Sirotin, Y. B. (2020). Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. PloS One, 15(8), e0237855–e0237855.

Jones, & Tranel, D. (2001). Severe Developmental Prosopagnosia in a Child With Superior Intellect. Journal of Clinical and Experimental Neuropsychology, 23(3), 265–273.

Kang, Y., & Harring, J. R. (2012). Investigating the impact of non-normality, effect size, and sample size on two-group comparison procedures: An empirical study. In *Annual Meeting of the American Educational Research Association (AERA)* (p. 29).

Kelly, Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The Other-Race Effect Develops during Infancy: Evidence of Perceptual Narrowing. Psychological Science, 18(12), 1084–1089.

Lee, & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies, 40(1), 153–184.

Levs, J. (2012, January 15). What caused the cruise ship disaster? CNN. Retrieved from https://www.cnn.com/2012/01/15/world/europe/italy-cruise-questions/index.html

Madhavan, Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. Human Factors, 48(2), 241–256.

Marsh, & Dibben, M. R. (2003). The role of trust in information science and technology. Annual Review of Information Science and Technology, 37(1), 465–498.

Meissner, & Brigham, J. C. (2001). THIRTY YEARS OF INVESTIGATING THE OWN-RACE BIAS IN MEMORY FOR FACES: A Meta-Analytic Review. Psychology, Public Policy, and Law, 7(1), 3–35.

Meyer. (2004). Conceptual Issues in the Study of Dynamic Hazard Warnings. Human Factors, 46(2), 196–204.

Noyes, E., & Hill, M. Q. (2021). Automatic recognition systems and human computer interaction in face matching. *Forensic face matching: Research and practice*, 193-215.

Osborne, J., & Overbay, A. (2008). Best practices in data cleaning. *Best practices in quantitative methods*, *1*(1), 205-213.

Parasuraman, & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 39(2), 230–253.

Parkes. (2013). Persuasive Decision Support: Improving Reliance on Decision Aids. Pacific Asia Journal of the Association for Information Systems, 4(3), 1–13.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). The use of multiple strategies in judgment and choice.

Ranjan, Bansal, A., Zheng, J., Xu, H., Gleason, J., Lu, B., Nanduri, A., Chen, J.-C., Castillo, C., & Chellappa, R. (2019). A Fast and Accurate System for Face Detection, Identification, and Verification. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(2), 82–96.

Rice. (2009). Examining Single- and Multiple-Process Theories of Trust in Automation. The Journal of General Psychology, 136(3), 303–322.

Ross. (2008). Moderators of trust and reliance across multiple decision aids. ProQuest Dissertations Publishing.

Russell, Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. Psychonomic Bulletin & Review, 16(2), 252–257.

Sawilowsky, & Blair, R. C. (1992). A More Realistic Look at the Robustness and Type II Error Properties of the t Test to Departures From Population Normality. Psychological Bulletin, 111(2),352-360.

Schwark, Dolgov, I., Graves, W., & Hor, D. (2010). The Influence of Perceived Task Difficulty and Importance on Automation Use. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54(19), 1503–1507.

Sengupta, S., Chen, J. C., Castillo, C., Patel, V. M., Chellappa, R., & Jacobs, D. W. (2016, March). Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-9). IEEE.

Simon. (1955). A Behavioral Model of Rational Choice. The Quarterly Journal of Economics, 69(1), 99–118.

Spain. (2009). The effects of automation expertise, system confidence, and image quality on trust, compliance, and performance. ProQuest Dissertations Publishing.

Susilo, Germine, L., & Duchaine, B. (2013). Face Recognition Ability Matures Late: Evidence From Individual Differences in Young Adults. Journal of Experimental Psychology. Human Perception and Performance, 39(5), 1212–1217.

Taigman, Ming Yang, Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 1701–1708.

White, Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. PloS One, 10(10), e0139827–e0139827.

White, Guilbert, D., Varela, V. P. L., Jenkins, R., & Burton, A. M. (2022). GFMT2: A psychometric measure of face matching ability. Behavior Research Methods, 54(1), 252–260.

Appendix 1

Red text shows changes from the questionnaires designed for Carragher, Sturman & Hancock

Initial Ability Estimates

1. How confident are you in your ability *to accurately judge whether two photographs show the same person*? (11 point scale from "extremely low" to "extremely high")
2. How accurate (as a %) do you think you will be when completing this task on your own (unassisted)? (0-100%)
3. Compared to the average person, do you think your unassisted performance will be above average or below average? (below average, above average)
4. How challenging do you think this task is going to be? (11 point scale from "extremely difficult" to "extremely easy")

Lee & Moray (1994)

Pre-Experience Trust Questions

The effect of trust in automation on human use of simulated Automated Facial

Ratings made on a scale from 0 to 10 ("extremely low" to "extremely high")

1. How confident are you in your ability *to accurately judge whether two photographs*

show the same person?

2. How much do you trust the facial recognition system *to accurately judge whether two photographs show the same person*? **Exploratory Questions**

3. Do you trust the facial recognition system to help you in this task? (no, yes)
4. How often do you think you will rely on (follow) the decisions made by the facial recognition system? (11 point scale: Never – Always)
5. The facial recognition system only makes errors on 5% of trials. How confident are you that you will be able to detect these errors? (11 point scale: Extremely Low – Extremely High)
6. If you disagree with a decision from the facial recognition system, how often do you think you will change your decision to match that given by the system? (e.g., if you think the two faces are the "same" person but the system says "different", how often will you change your mind and report "different" too?) (11 point scale: Never – Always)
7. To what extent do you believe the facial recognition system's decisions will influence your performance on the task? (11 point scale: 'significantly impair' to 'significantly improve')
8. To what extent do you believe that the facial recognition system's decisions will influence the difficulty of the task? (11 point scale: 'significantly harder' to 'significantly easier')
9. Who do you think will be more accurate at this task? (You, The Facial Recognition System)

10. If you could choose your source of help for this task, would you prefer to see the decisions made by a facial recognition system or another person? (Another Human/ A Facial Recognition System)
11. Outside of this experiment, how would you rate your level of knowledge about facial recognition technology? (11 point scale: “Extremely low” to “Extremely high”)
12. How accurate (as a %) do you think you will be when completing this task on your own (unassisted)? (0-100)
13. Compared to the average person, do you think your unassisted performance will be above average or below average? (below, above)
14. How accurate (as a %) do you think the facial recognition system will be on this task? (0-100)
15. How accurate (as a %) do you think you will be if you complete this task with the assistance of the facial recognition system? (0-100)
16. How challenging do you think this task is going to be? (11 point scale from “extremely difficult” to “extremely easy”)

****

Post-Experience questions

Lee & Moray (1994)

Ratings made on a scale from 0 to 10 (“extremely low” to “extremely high”)

1. How confident are you in your ability *to accurately judge whether two photographs show the same person*?
2. How much do you trust the facial recognition system *to accurately judge whether*

two photographs show the same person?

Exploratory Questions

3. Do you trust that the facial recognition system helped you in this task? (no, yes)

4. How often do you think you relied on (followed) the decisions made by the facial recognition system? (11 point scale: Never – Always)
5. The facial recognition system only made errors on 5% of trials. How confident are you that you were able to detect these errors? (11 point scale: Extremely Low – Extremely High)
6. If you disagreed with a decision from the facial recognition system, how often do you think you changed your decision to match that given by the system? (e.g., if you thought the two faces were the "same" person but the system said "different", how often did you change your mind and report "different" too?) (11 point scale: Never – Always)
7. To what extent do you believe the facial recognition system's decisions influenced your performance on the task? (11 point scale: 'significantly impair' to 'significantly improve')
8. To what extent do you believe that the facial recognition system's decisions influenced the difficulty of the task? (11 point scale: 'significantly harder' to 'significantly easier')
9. Who do you think was more accurate at this task? (You, The Facial Recognition System)
10. If you could choose your source of help for this task, would you prefer to see the decisions made by a facial recognition system or another person? (Another Human/ A Facial Recognition System)

11. How accurate (as a %) do you think you would have been without the assistance of the facial recognition system? (0-100)
12. Compared to the average person, do you think your unassisted performance (first decisions) was above average or below average? (below, above)
13. According to the instructions, how accurate was the facial recognition system? (55%, 95%)
14. The instructions said that the facial recognition system would give the correct answer on 95% of trials. Do you believe that the facial recognition system was that accurate? (no, yes)

    IF NO, show 15

15. How accurate (as a %) do you think the facial recognition system was on this task? (0-100)

16. How accurate (as a %) do you think you were when completing this task with the assistance of the facial recognition system? (0-100)

17. How challenging did you find this task? (11 point scale from “extremely difficult” to “extremely easy”)