Title Page

The Effect of Perceived Task Role on Reliance on Artificial Intelligence in Unfamiliar Face Matching Tasks

a1771647

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

School of Psychology, University of Adelaide, Australia

Word Count: 6083

Table of Contents

List of Figures

List of Tables

Abstract

Unfamiliar face matching is the process of observing two faces and determining whether they belong to the same person (a match) or two different people (a mismatch). Primarily required in security and identification contexts, this task is surprisingly difficult for humans. With Artificial Intelligence becoming increasingly powerful in automating mundane tasks, current state-of-the-art Automated Facial Recognition Systems (AFRS) can greatly outperform their human counterpart; however, they still often require human supervision and/or input. The 'human-machine interaction' is a term that describes the way humans and machines, in this case AFRS, function together. Whilst the impact of factors such as perceived responsibility and self-reliance on behaviour has been observed with respect to between-human interactions, their effect on the human-machine interaction remains mostly unexplored. This study aims to explore whether manipulating the perceived role in the human-machine interaction can affect trust in automation, complacency, automation-reliance, and ultimately performance in an AFRS-assisted unfamiliar face matching task. Whilst we observed a clear increase in performance when AFRS-assistance was introduced, we found no significant change in performance or trust based on perceived role. Furthermore, human operators curtail the performance of the AFRS regardless of their perceived role in the human-machine interaction.

Keywords: Face Matching, Perceived Role, Trust in Automation, Human-Machine Interaction

Declaration

Contributor Roles Table

ROLE	ROLE DESCRIPTION	STUDENT	SUPERVISOR 1
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
SUPERVISION	Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.		X
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	X
FORMAL ANALYSIS	Application of statistical, mathematical, computational, or other formal techniques to analyze 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

Introduction

Unfamiliar face matching is the process of determining whether two faces belong to the same person (denoting a 'match') or two different people (denoting a 'mismatch'; Alenzi & Bindemann, 2013). The need for unfamiliar face matching commonly arises in scenarios concerned with identification and security, such as passport control and ID verification (Fysh & Bindemann, 2017), and typically takes the form of either image-to-image matching (in which two photographs are compared) or face-to-image matching (in which a photograph is compared to a live person). While the majority of real-world instances of face matching involve face-to-image matching, image-to-image matching is also utilised in specific situations such as identity database searches and camera footage face comparison (Moreton, 2021). Note that it is important to distinguish between face matching and face recognition tasks, the latter referring to observing a single face and determining whether it has been seen before (Bruce & Young, 1986). This type of task, typically used in police line-ups, eye-witness testimony, and a multitude of other forensic applications, is a fundamentally different task to face-matching in that it is a task of memory rather than perception (Lamont et al., 2005).

Humans are surprisingly bad at simple unfamiliar face matching tasks (White et al., 2014). The average accuracy for humans in image-to-image face matching varies depending on the task, but is commonly around 80% (Burton et al., 2010), with face-to-image task accuracy being reportedly as low as 67% (Kemp, Towell & Pike, 1997). Although average performance is low, the task of unfamiliar face matching exhibits extreme between-subject variation; with overperformers being referred to as "super recognisers". For example, in one such study where the mean accuracy of control participants was 73.6%, a group of "super recognisers" had a mean accuracy of 90.3% (Robertson et al., 2016). While the existence of super recognisers as a distinct subset of people is contested (Noyes, Phillips & O'Toole, 2017), the concept is useful as a means of highlighting the extreme discrepancy in face matching performance between individuals (Bobak, Hancock & Bate, 2016). Interestingly, unfamiliar face-matching ability appears to be stable over time, and performance has been observed to be unrelated to amount of occupational experience (Robertson et al., 2016; White et al., 2014).

Artificial intelligence in Unfamiliar Face Matching Tasks

Many scenarios in which unfamiliar face matching tasks are required have implemented Automated Facial Recognition Systems (AFRS) in conjunction with human operators (Towler et al., 2017). Unlike humans, AFRS are capable of processing extremely large amounts of data in a small timeframe, and are unaffected by biological and psychological factors such as fatigue and cognitive variability (Alenezi et al., 2015). The way AFRS operate is exceedingly complex, but a simplification of the process is as follows: they first analyse the presented images and attempt to identify the presence of a face, they then locate and assess relevant 'features' of each image (the means by which this is accomplished varies between algorithms), which are then compared between images to yield a 'similarity score'; and finally, this score is compared to a threshold value to determine whether the faces are a match (similarity score above threshold) or a mismatch (similarity score below threshold; Noyes & Hill, 2021).

Importantly, many modern algorithms tend to perform significantly better than average humans at basic unfamiliar face matching tasks (Carragher & Hancock, 2022). For example, where one study (Stacchi et al, 2020) reported that the average accuracy for human participants (n = 181) on the Expertise in Face Matching Test (White et al., 2015) was 77.9% (SD = 8.2), a modern AFRS (a Deep Convolution Neural Network; DCNN) correctly identified all 168 face pairs (Carragher & Hancock, 2022). The use of AFRS in an applied face matching setting typically consists of one or more human operators conducting oversight of an AFRS, watching for and correcting its errors (Fysh & Bindemann, 2018). This type of implementation is referred to as a 'human-machine interaction.'

Human-Machine Interaction in Face Matching

Human-machine interaction is defined as "the interaction and communication between human users and a machine, a dynamic technical system, via human-machine interface" (Johanssen, 2009, p. 132). Whilst there exists a plethora of research regarding the human-machine interaction in general (Carrol, 1997; Hancock et al., 2011), there is little research regarding the human-machine interaction in the context of face matching. Fysh & Bindemann (2018) were responsible for the foundational study of human-machine interaction in face matching, in which participants were presented with a

simple unfamiliar face matching task (as described earlier) with each trial accompanied by a decision from an AFRS ('same', 'different', or 'unresolved'). By reducing the accuracy of the AFRS, participants exhibited significant decrease in performance, showing that human operators relied heavily on the AFRS when completing face matching tasks.

A following study by Howard et al. (2020) consisted of a similar exercise, but included a control group (with no assistance) as well as a 'human-aided' group, and also incorporated confidence ratings for each trial. They observed an effect in which participants reported trusting human assistants less than automated ones, but did not show any difference in assisted performance. However, it is to be noted that this study subjected each participant to only 14 face matching trials, and unassisted performance matched AFRS-aided performance, thus the validity of their findings is questionable. The current study borrows from the most recent study of human-machine interaction in face matching (Carragher & Hancock, 2022), in which a series of AFRS of differing accuracy and format were compared to one another and unassisted participants. Among these previous studies, an aspect of the human-machine interaction in face matching that remains insufficiently explored is the role of trust in automation.

Trust in Automation and Human-Machine Interaction

Unsurprisingly, the extent to which a human operator trusts the ability of a machine has important consequences on the outcome of a human-machine interaction (Mayer et al., 2023), such that if the human member of the team does not trust the machine to contribute meaningfully or consistently to the task, they are less likely to rely on it. This act of neglecting to utilise a machine is referred to as 'disuse', whilst 'misuse' refers to negligent overreliance on a machine (Parasuraman & Riley, 1997). In cases where machines are more proficient than humans, their disuse often results in worse performance in tasks involving human-machine interaction (Lyons & Guznov, 2019). Although it has been shown that one's trust in a machine is correlated with the one's perception of said machine's reliability (Moray, Inagaki & Itoh, 2000), Dzindolet et al. (2002) observed a phenomenon in which participants disused a reliable automated aid – an effect they attributed to self-reliance (a proneness to feeling responsible to complete something by oneself).

Conversely, there are examples of consistent machine misuse in certain scenarios due to human complacency in human-machine interactions (Parasuraman & Manzey, 2010). Complacency in this context denotes *"*self-satisfaction that may result in non-vigilance based on an unjustified assumption of satisfactory system state*"* as defined in the NASA Aviation Safety Reporting System (Billings et al., 1976, p. 23). Whilst complacency has been shown to correlate negatively with situational awareness, attention and trust, and positively with task complexity (Molloy & Parasuraman, 1996; Parasuraman et al., 2008), it appears as though lack of attention is the strongest predictor of complacency (Parasuraman et al., 2010). Complacency emerges as a difficult issue to solve in situations where machines greatly outperform humans. For example, if a machine is to achieve 95% accuracy on a task where a human would score 70%, and the human were to simply allow the machine to do everything, they'd very likely score higher than if they contributed in any way. Thus, if we were to look at raw performance score to assess the quality of the interaction, those who are most complacent, and therefore contribute the least, would by accuracy alone be considered the 'best' users of the machine (Carragher & Hancock, 2022). This interaction would fail to achieve its intended goal; for the human operator to amend AFRS errors.

Perceived role

Noyes and Hill (2021) outline how rapid advancements in AFRS technology have led to a shift in dynamic in the human-machine interaction; such that whilst human operators once completed face matching tasks with AFRS assistance, it is now optimal for the more proficient AFRS to complete the task with the human operator performing oversight. Furthermore, Chiou & Lee (2023) suggest in their narrative literature review that the structure of a human-machine interaction may affect trust in automation and consequently performance. However, the effects of this change of role remain to be explored in the context of face matching. Since there appear to be no other examples of research into perceived role in the human-machine interaction, literature regarding human-to-human interaction was used to guide some theoretical aspects of this study.

Recent research into social factors and their role in perpetuating the ideas of self-reliance suggests that people may be less likely to accept help when it comes from peers/subordinates as

opposed to superiors (Thompson & Bolino, 2018). This notion is further supported by Schyns et al. (2022), who suggest both individual differences and social dynamics can affect one's inclination to lead and/or accept assistance. Additionally, there is evidence to suggest that perceived social status is positively correlated with confidence, and in turn, likelihood to take initiative (Kim et al., 2022). The implication of this is that if these findings apply to the human-machine interaction, then manipulating a human operators' perception of their role in a human-machine interaction could affect the extent to which they misuse/disuse a machine. It is important to note that these examples involve cross-sectional studies of complex work environments, and thus may not be directly applicable to human-machine interaction; however, they are among the few papers currently exploring perceived role's effects on human behaviour. Thus, it is the goal of this study to determine whether these findings can be generalised to human-machine interaction.

The Present Research

The primary aim of this study is to explore whether a human operator's perceived role in a human-machine interaction affects performance in an unfamiliar face matching task. Specifically, we will investigate whether participants who are instructed to complete the task whilst "assisted" by an AFRS will perform differently on the face matching task to those who are instructed to perform "oversight" over the AFRS's answers.

We hypothesise that participants in the oversight condition will be more trusting of the AFRS than participants in the assisted condition. Furthermore, since the AFRS will likely perform better than participants, and we expect those who trust the AFRS more to accept its input more, we hypothesise that participants in the oversight group will achieve higher accuracy than those in the assisted group when working with the AFRS. Additionally, we expect that participants in the assisted group will be more prone to self-reliance, thus being more likely to overturn AFRS decisions. Conversely, we expect participants in the oversight condition to be more complacent, and thereby more likely to misuse the AFRS than the assisted condition, resulting in them correcting fewer of the AFRS' errors.

Method

Sample Size

An a priori power analysis was performed using G*Power (Faul et al., 2007) to determine the required sample size for the study. Given a conventional alpha of .05 and a medium effect size of η_p^2 = .06, a sample size of 126 participants was required to detect an effect with 80% power in a 2 x 2 mixed measures ANOVA.

Participants

Participants were first year undergraduate psychology students from the University of Adelaide who completed the study for course credit. The mean age of participants was 19.18 (*SD* = 2.04). A final sample size of 73 participants was achieved after excluding those who failed the attention check trial (*n* = 3), those who failed to recall the accuracy of the AFRS (*n* = 3), those who took longer than 60 minutes to complete the experiment (*n* = 3), those who attempted the experiment multiple times (*n* = 3), and one participant who failed to finish the experiment. The sample of 73 participants was below the recommended sample size as per the a priori power analysis. The study was approved by the Human Resource Ethics Sub-Committee of the University of Adelaide (H-2019-23/01), and informed consent was obtained from every participant.

Design

This study utilises experimental design, with participants randomly allocated to either the 'oversight' or 'assisted group'. The experiment is split into two phases; the baseline phase and the aided phase. Thus, there is both a 'perceived role' condition as well as a 'task phase' condition.

Materials

Glasgow Face Matching Test 2 short version (GFMT-2S)

The GFMT-2S (White et al., 2022) is a measure of unfamiliar face matching ability. It consists of 80 pairs of faces split into two equally difficult groups of 40 pairs. In each group, 20 pairs are a 'match' such that they belong to the same person, and 20 pairs are a 'mismatch' such that they

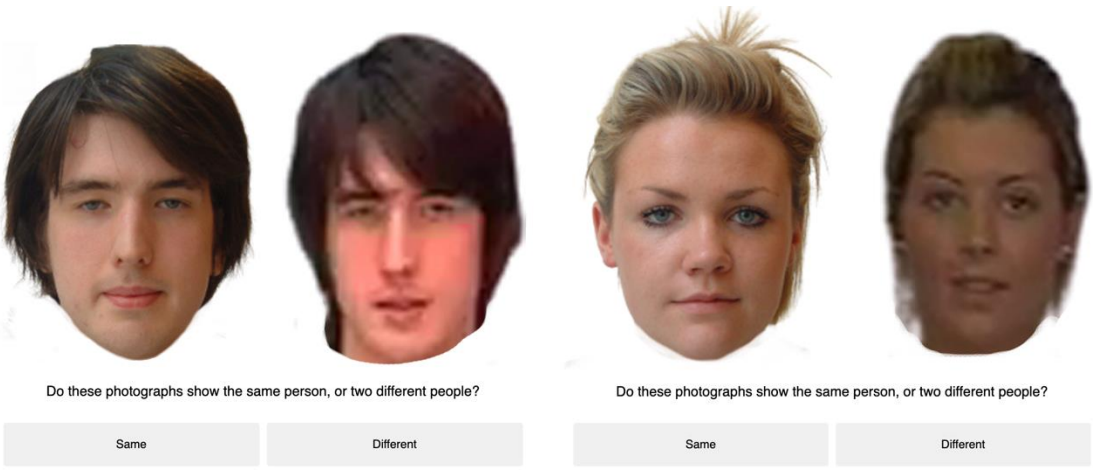
belong to different people. The GMFT-2S was chosen for its length, high test-retest reliability, $r(107) = .774$, and high internal consistency (Cronbach's alpha = .938).

Each phase of the experiment includes 40 trials of the GFMT-2S, which are presented as the pair of faces alongside two buttons labelled 'same' and 'different' (see Figure 1). For each trial, the participant must respond either 'same', indicating that the faces were a match or 'different' indicating that they were a mismatch. After each trial, participants provided a confidence score for their decision using a scale ranging from 1 (not confident at all) to 7 (extremely confident). During the aided phase, each trial is accompanied by the AFRS' decision (see Figure 2).

Figure 1

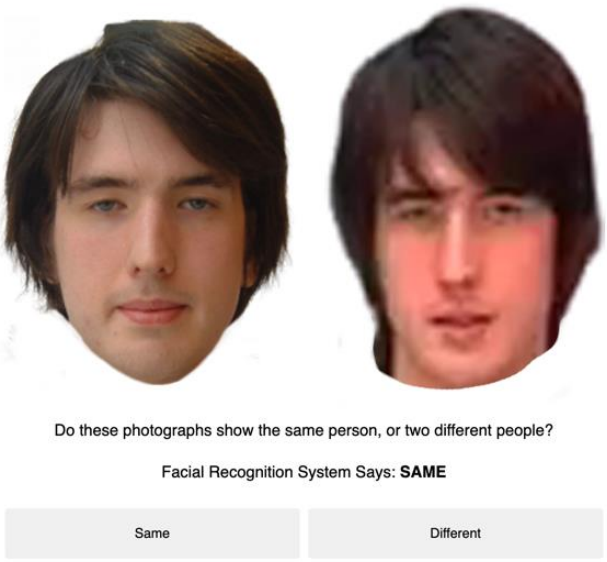
Example of a Match and Mismatch Trial in the Baseline Phase

a) b)



Note: An example of a 'match' pair (a) and a 'mismatch' pair (b) as they are presented in the experiment during the baseline phase.

Figure 2
Example of a Match Trial in the Aided Phase



Note: An example of a 'match' pair alongside the AFRS's decision (below the question). In this instance the AFRS is correct. This example trial is representative of the aided phase for both the 'assisted' and 'oversight' conditions.

Manipulation

After the baseline phase but before the aided phase, each participant was briefed with instructions for the aided phase, with participants receiving different instructions based on their allocated group (note that this difference in aided phase instructions was the sole manipulation of this experiment). The instructions for each group were as follows:

Oversight Instructions. "Your job is to oversee the identification decisions made by the face recognition system regarding whether pairs of faces are identity matches (i.e., the same person) or mismatches (i.e., two different people). You will see that the system has made a "SAME" or "DIFFERENT" decision for each pair. The decisions from the face recognition system will be correct 95% of the time. You will be responsible for manually entering the answer on each trial. As such, you can input the face recognition system's decision on most trials. However, you should change the decision from the face recognition system when you believe that the system has made an error."

Assisted Instructions. "Your job is to decide whether pairs of faces are identity matches (i.e., the same person) or mismatches (i.e., two different people). On each trial, you will see the identification decision made by the face recognition system. Like you, the system will give a "SAME"

or "DIFFERENT" decision for each pair. The decisions from the face recognition system will be correct 95% of the time. You will be responsible for making the final identification decision for each pair, but you can use the facial recognition system's answer to help make your choice. Your goal is to try to achieve 100% accuracy with the assistance of the face recognition system."

Simulated Automated Facial Recognition System (AFRS)

The simulated AFRS is modelled after a real artificial intelligence's (Deep Convolutional Neural Network) performance on the GFMT-2S (see Carragher & Hancock, 2022). Additional errors were manually added into the 80 trials, changing the AFRS' accuracy from 97.5% to 95% to allow participants more opportunities to correct its mistakes. The simulated AFRS is represented by a prompt on each face matching trial that indicates whether the system has concluded that the faces are the 'same' or 'different' (see Figure 2).

Attention checks

Trial Attention Check. The first attention check consists of a pair of famous faces that is a clear mismatch as they are clearly of different ethnicity or gender. This attention check is implemented in each phase of the experiment (baseline, aided).

Post-Experiment Question. The second attention check consists of a multiple-choice question after the aided phase of the experiment in which each participant must correctly identify the accuracy of the AFRS as is detailed in the task instructions.

Participants who incorrectly identified either of the famous mismatch pairs as 'same' or incorrectly identified the AFRS's accuracy had their data excluded from the analysis.

Measures

Trust/Self Confidence Measure

The Trust/Self Confidence Measure (Lee & Moray, 1994) is a self-report questionnaire consisting of two Likert-type items ranging from 1 (strongly disagree) to 10 (strongly agree). One question concerns the participant's confidence in their ability to complete the task, while the other question concerns the participant's trust in the automated system to complete the task. The confidence score is subtracted from the trust score resulting in a value that represents a participants' trust in automation relative to their trust in their own ability.

Exploratory Trust in AFRS

The trust in AFRS measure is a set of twelve exploratory self-report questions developed by Carragher, Sturman & Hancock (2023). These questions cover a variety of aspects of trust in the form of either Likert-type or multiple choice, some examples are: "Do you trust the facial recognition system to accurately decide whether two photographs show the same person?", "How often do you think you will agree with the decisions made by the facial recognition system?", "How accurate (as a %) do you think you will be when completing this task on your own (unassisted)?". These questions are used in exploratory analyses comparing various aspects of trust with performance and between groups.

NASA Task Load Index (NASA TLX)

The Task Load Index (Hart, 1986) is a self-report questionnaire consisting of 6 likert-type items ranging from 1 (very low) to 7 (very high) regarding perceived task stress and/or effort. All items are positively coded, such that an answer of 'very high' always indicates high task load. It is noted that although item two – which concerns physical demand – is irrelevant to the entirely cognitive task, it was included as to not affect interpretations of the final score. The TLX has high internal consistency (Cronbach's alpha = .86), as well has high concurrent and structure validity (Xiao et al., 2005). It was included as a manipulation check to measure if there was difference in effort between groups.

Stirling Face Recognition Scale (SFRS)

The SFRS (Bobak, Mileva & Hancock, 2019) is a self report questionnaire consisting of 20 questions related to one's perception of their own face recognition ability. All questions are likert-type items ranging from 1 (strongly disagree) to 7 (strongly agree). The SFRS has high reported internal consistency (Cronbach's $\alpha = .88$), and is used to measure each participant's self-perceived face recognition ability.

Procedure

The study was run entirely within Qualtrics, an online survey program. Participants were first introduced to the study and informed consent was obtained. Following this, each participant was instructed about the task, and completed an initial accuracy estimate and the SFRS to determine self-perceived face recognition ability. Each participant then completed the 40 baseline trials. After the baseline phase, participants completed the baseline NASA TLX and were then introduced to the AFRS and instructed on the second phase of the study. In these instructions, participants in the assisted group were to complete the task with the assistance of the AFRS, whereas participants in the oversight group were told to enter the AFRS' decisions whilst correction any errors. Each group then completed the Trust in AFRS questionnaire, followed by 40 aided trials**.** Importantly, the aided phase of the experiment was identical for each group, with the discrepancy in instructions being the only manipulation. After completing the aided phase, participants completed the post-experiment section of the Trust in AFRS questionnaire, followed by the post-experiment NASA TLX before being debriefed and dismissed.

Analysis

All measures used in the primary analysis were assessed and confirmed for homogeneity of variance, and all but two measures (aided phase accuracy, SFRS scores) were normally distributed as assessed via Shapiro-Wilk test of normality. Although these two variables violated the assumption of normality associated with ANOVAs and ANCOVAs, as there is no non-parametric alternative to an ANCOVA, we continued as planned. Complacency was assessed via NASA TLX scores, number of

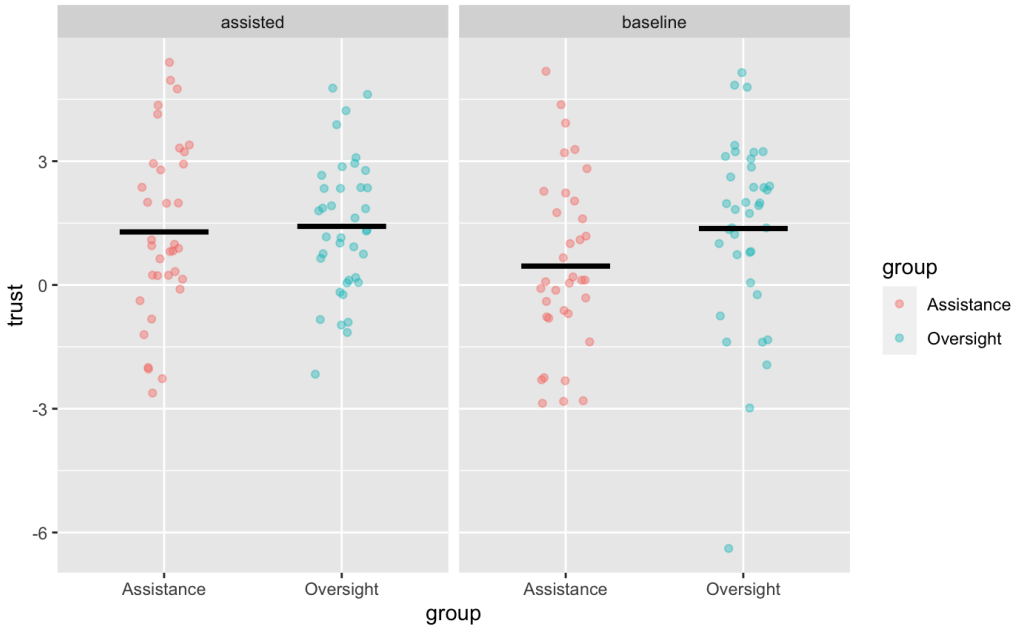
AFRS errors overlooked, and reported perceived responsibility. All data analysis was conducted using R and RStudio version 2022.07.1.

Results

Primary Analysis 1: The Effect of Perceived Role on Trust

To see if perceived role had an effect on Relative Trust in Automation, we conducted a 2 (Perceived Role: Assisted, Oversight) x 2 (Phase: Baseline, Aided) mixed ANOVA with Relative Trust in Automation as the dependent variable. We found no significant difference in Relative Trust in Automation between the Oversight group and the assisted group, $F(1, 71) = 1.64$, $p = .204$, $\eta_p^2 = .023$; nor was there a significant phase effect, $F(1, 71) = 3.50$, $p = .065$, $\eta_p^2 = .047$, or group x phase interaction $F(1, 71) = 2.72$, $p = .104$, $\eta_p^2 = .037$ (See Figure 3).

Figure 3
Trust scores pre- and post-aided phase for each group



Note: each circle represents one participant, the back bar represents mean accuracy

Primary Analysis 2: The Effects of Perceived Role on Performance

Descriptive statistics of all measures relevant to the primary analyses are shown in Table 1. The first primary analysis consists of a 2 (Perceived Role: Assisted, Oversight) x 2 (Phase: Baseline,

Aided) mixed ANOVA with face matching accuracy as the dependent variable. We found a significant main effect of phase, such that mean accuracy was higher in the aided phase (M = 90.58, SD = 6.90) than in the baseline phase (M = 86.06, SD = 7.53), $F(1, 71) = 20.52$, $p < .001$, $\eta_p^2 = .224$. We found no difference in improvement over baseline between the assisted group and the oversight group $F(1, 71) = 0.42$, $p = .519$, $\eta_p^2 < .01$, indicating that perceived role had no significant effect on accuracy. Similarly, the interaction effect between perceived role and phase was not statistically significant $F(1, 71) = 1.80$, $p = .18$, $\eta_p^2 = .02$. These findings indicate that while participants performed better during the aided phase of the experiment, the oversight group did not perform better than the assisted group, nor did they improve more in the aided phase (see Figure 4).

Table 1

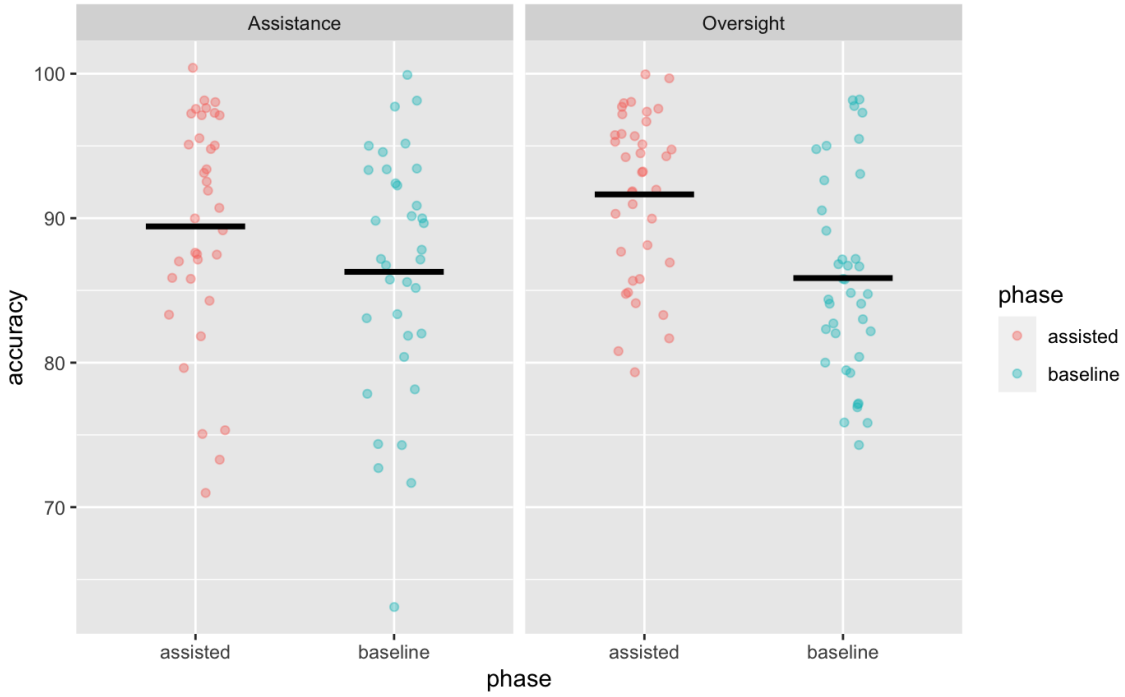
Means, Standard Deviations, and Differences in Means for Measures used in Primary Analyses by perceived role

	Assist			Oversight			Difference
	N	**Mean**	**SD**	**N**	**Mean**	**SD**	**t**
Overall Accuracy	35	87.86	6.563	38	88.75	5.174	0.64
Baseline Accuracy	35	86.29	8.388	38	85.86	6.76	-0.24
Aided Accuracy	35	89.43	7.931	38	91.64	5.703	0.13
Accuracy Change	35	3.143	9.708	38	5.789	7.026	1.32
SFRS Score	35	99.23	17.45	38	101.8	16.55	0.65
Baseline TLX Score	35	20.91	5.21	38	20.24	5.572	-0.54
Aided TLX Score	35	20.23	6.236	38	20.05	5.291	-0.13
Baseline TiA Score	35	0.4571	2.063	38	1.368	2.186	1.82*
Aided TiA Score	35	1.286	2.094	38	1.421	1.671	0.30

Notes: Statistical significance markers: * p<0.1; ** p<0.05; *** p<0.01, *Accuracy Change = difference in accuracy between baseline phase and aided phase, Baseline TiA Score = Trust in Automation score before phase 2, Aided TiA Score = Trust in Automation score after phase 2.*

Figure 4

Accuracy scores pre- and post- aided phase by perceived role



Primary Analysis 3: The Interaction between Trust in Automation, Self-Perceived Ability, Skill, and Role on Performance

The third primary analysis consists of a 2 (Perceived Role: Assisted, Oversight) x 2 (Phase: Baseline, Aided) mixed ANCOVA with baseline accuracy, trust in automation, and self-perceived ability as covariates. Unsurprisingly, baseline performance was the strongest predictor of aided performance, $F(1, 20) = 172.03$, $p < .001$, $\eta_p^2 = .90$. Relative trust in automation had no significant effect on accuracy $F(1, 20) = 1.21$, $p = .285$, $\eta_p^2 = .05$, and neither did self-perceived ability, $F(1, 20) = 3.81$, $p = .065$, $\eta_p^2 = .16$. The group effect remained statistically non-significant once covariates were accounted for, $F(1, 20) = 0.32$, $p = .578$, $\eta_p^2 = .02$. These findings indicate that neither self-perceived task confidence nor trust in automation had an effect on overall performance nor did they interact with the group or phase effects.

Primary Analysis 4: The Effect of Perceived Role on Complacency

To determine whether role had an effect on complacency, a series of t-tests were run comparing mean TLX scores, the number of AFRS errors overturned, and reported perceived responsibility between groups. There was no significant difference in number of errors overturned between the Oversight group and the Assisted group, $t(71) = 0.36$, $p = .723$. Similarly, participants in

the Oversight group did not report lower perceived responsibility than the Assisted group, $t(71) = -0.33$, $p = .743$. Lastly, there was no significant difference in aided-phase TLX scores between the Oversight and Assisted conditions, $t(71) = .13$, $p = .897$, indicating that those in the oversight group did not report lower expended effort.

Additionally, we ran a 2 (Perceived Role: Assisted, Oversight) x 2 (Phase: Baseline, Aided) mixed ANOVA comparing the number of responses participants submitted that were the same as the AFRS'. This was done in order to determine whether those in the oversight group were more likely to follow the AFRS' decision than those in the assisted group. We found no difference in response similarity between groups, $F(1, 71) = .38$, $p = .541$, $\eta_p^2 = .005$, nor was there a group x phase interaction, $F(1, 71) = .83$, $p = .365$, $\eta_p^2 = .012$. Overall, these findings indicate that there was no difference in complacency between groups.

Exploratory Analyses

The Relationship Between Trust in Automation and Improvement

To see if trust in automation correlated positively with improvement, we conducted multiple Pearson's correlations between participants' baseline and aided relative trust in automation scores and their accuracy improvement over baseline. There was no correlation found between pre-AI-integration trust scores and improvement, $r(71) = 0.04$, $p = .761$, or between post-AI-Integration trust scores and improvement, $r(71) = -.08$, $p = .519$.

Replicating Carragher & Hancock's (2022) Findings that Aided Performance was Worse than the AFRS Alone

To see if the human-AFRS teams resulted in better performance than the AFRS alone, we ran two one-sample t-tests comparing mean aided-phase accuracy of participants in the oversight group and the AFRS, and participants in the assisted group and the AFRS respectively. Aided participants achieved significantly lower accuracy than the AFRS alone in both the oversight group, $t(37) = -3.63$, $p = <.001$, and the assisted group, $t(34) = -4.16$, $p = <.001$.

Replicating Howard et al.'s (2020) Per-Trial Confidence Increase

A 2 (Perceived Role: Assisted, Oversight) x 2 (Phase: Baseline, Aided) mixed ANOVA with average trial confidence as the dependent variable was run to see if we could replicate the effect observed in Howard et al.'s (2020) study, in which participants reported significantly higher decision confidence when aided than when performing the task alone. We found a main effect of phase such that participants were significantly more confident with their decision in aided trials (M = 5.34, SD = 0.73) than in baseline (unaided) trials (M = 5.01, SD = 0.67), $F(1, 71) = 43.35$, $p < .001$, $\eta_p^2 = .400$, and there was no significant group x phase interaction, $F(1, 71) = 1.07$, $p = .306$, $\eta_p^2 = .015$.

Discussion

This study compared unfamiliar face matching performance, trust in automation, and complacency between participants who were told to perform oversight of an AFRS and those who were told to use the AFRS as an aid in an unfamiliar face matching task. We found no significant difference in face matching accuracy, relative trust in automation, or complacency between participants in the oversight and aided groups. We did however observe a strong phase effect, such that regardless of group, participants achieved significantly higher accuracy when working with the AFRS than when performing the task alone. However, the majority of participants still failed to match or exceed the accuracy of the AFRS by itself. In our exploratory analyses, we found that the Stirling Face Recognition Scale (Bobak, Mileva & Hancock, 2019) had no correlation with face matching performance, and that participants reported higher decision confidence when working with the AFRS than when alone.

Compared to previous findings, we noticed some consistencies as well as discrepancies. The observed increase in task performance due to AI-assistance, as well as the tendency for participants to curtail the performance of the AFRS, were consistent with the findings of both Carragher & Hancock (2022) and Fysh & Bindemann (2018). Additionally, we observed extremely high population variance in face matching ability as well as extremely high levels of stability in face matching ability; as is consistent with Bobak, Hancock & Bate (2016) and Robertson et al. (2016) respectively. Furthermore, most participants who achieved an aided accuracy higher than that of the AFRS alone had extremely

high accuracy in the unaided phase. These findings suggest that a human-AFRS team requires a highly proficient human face matcher to function optimally.

However, we also found that baseline relative trust in automation had no correlation with improvement, despite Carragher, Sturman and Hancock (2023) finding a significant positive correlation. These results are however consistent with Howard et al.'s (2020) findings, in which participants' trust of a face matching aid did not correlate with aided performance. Similarly, we observed that participants' per-trial confidence scores increased significantly in the aided phase; an effect also consistent with Howard et al.'s (2020) findings. Interestingly, the method of this study more closely matches that of Carragher, Sturman and Hancock (2023) as opposed to Howard et al (2020), and as such we would expect these results to replicate the former. There are several reasons that could explain the lack of this effect. Firstly, the average baseline performance of all participants was noticeably higher in our study (M = 86.06) than in Carragher, Sturman and Hancock's (M = *82.23)*, leaving significantly less room for improvement in the aided phase. Furthermore, Carragher, Sturman and Hancock's (2023) effect was small among a sample of 174 participants, a finding our study likely did not have the statistical power to detect.

Another finding we were unable replicate is that of Parasuraman et al. (2008), in which they suggest trust in automation was positively correlated with complacency; however, we acknowledge that there are potential reasons as to why we could not replicate this effect. For one, Parasuraman et al. (1996) found that complacency is related to both task length and complexity. Given that the task in our study was short, simple, and to be performed without distraction, it is possible that the task was neither long enough nor complex enough for participants to need to rely on the AI; thus, the environment may not have been appropriate for complacency to eventuate.

One interesting thing to note about the current study's results is that although the group effect of relative trust in automation was not statistically significant, there appeared to be a pattern in which those in the oversight condition trusted the AFRS more based on the task instructions alone. However, once all participants had used the AFRS, post-aided phase trust scores were near identical between groups. This finding is consistent with both Madhaven et al. (2006) and Ross et al. (2008), in which it is shown that people's trust in automation is related to the reliability of the automation. Since both

groups used an identically reliable AFRS, their post-aided phase trust in the AFRS converged. An inference of this is that it may be the case that despite attempted manipulation of trust in automation, it will tend towards a level appropriate to the actual reliability of the machine. This could have potentially been a factor in nullifying our expected effect in which perceived role influenced trust.

Strengths and Limitations

This study was the first of its kind in attempting to manipulate perceived role in the human-machine interaction without simultaneously altering the nature of the task completed. As such, in creating our manipulation, we relied heavily on concepts untested in the fields of face matching and human-machine interaction. Thus, a potential shortcoming of the study would be that the manipulation may not have suitable and/or substantial in the context of human-machine interaction. This notion is supported by the absence of any significant effect in our manipulation checks. Since the oversight group failed to report lower task load and/or higher trust in the AFRS – the factors related to misuse – it is logical that we would see no increase in misuse for this group. The inverse is also true of the assisted group, such that since we did not observe lower trust in automation or higher task load, we can reasonably infer that the assisted group did not exhibit a self-reliance effect leading to disuse. These findings indicate that it may not be the case that perceived role does not influence these factors, but rather that we failed to significantly influence participants' perceived role in the experiment.

Another limitation of this study worth noting is the failure to meet an appropriate sample size for our analyses. This is especially relevant given that the expected effect sizes of our interactions were small. The lack of an interaction between perceived role and phase, however, does not appear to be attributable to insufficient detection power; as although it is true the study was underpowered, the effect size of group on accuracy was so low (<.01) that even with an appropriate sample size it would be unlikely to become significant. This also appears to be the case with the difference in number of errors overturned between groups.

It is important to note that this study was not intended to replicate a realistic environment in which an AFRS would be used in conjunction with a human operator. One of the main discrepancies between which would be the number of mismatched trials as a percentage of total trials. We elected to

include a high number of mismatches for several reason. Firstly, if the amount of mismatches represented reality, then given the length of the task, the number of mismatches would need to be close to, if not, zero. Assuming the number of mismatch trials was made to be low but not realistically so (say 10%), a different problem would emerge in which the task becomes too easy due to match trials being significantly easier than mismatch trials (Fysh & Bindemann, 2018), causing accuracy scores to be too close to the upper limit in the baseline phase for meaningful improvement in the aided phase to be possible. Additionally, by using a face matching task consistent with previous studies (Carragher & Hancock, 2022; Howard et al., 2020), we can make more valid comparisons between findings.

Other aspects in which the experiment differs from reality are in its environment, difficulty and length of the task. Whereas real AFRS operators tasked with face matching would perform thousands of instances across many hours in a distraction-riddled environment, our experiment was fast, of moderate difficulty, and performed at the participants volition. Although this may be an attributing factor to our lack of observed complacency and/or AFRS misuse, it is not strictly a shortcoming of the study. Whilst it may be true that complacency is heavily influenced by attention as well as task length and difficulty (Parasuraman et al, 1997), by conducting this experiment with those contributing factors absent, we can observe the effects of perceived role on complacency in a near-vacuum. Thus, if a complacency effect were to emerge, it would indicate that complacency can eventuate in a task without it being difficult or lengthy. Lastly, it is worth mentioning that although the sample of this study is not strictly representative of the target population (those who perform unfamiliar face-matching occupationally), it has been shown that occupational face matchers are, on average, no better than amateurs (Robertson et al., 2016).

Future directions and Concluding Thoughts

Further research regarding the effects of perceived role in the human-machine interaction in unfamiliar face matching can expand upon several aspects of our study. Potential iterations of this study with multiple levels of manipulation may be worth exploring. For example, a condition in which the participant is told they are solely responsible for the final decision vs. The AFRS is solely responsible for the final decision. However, by increasing the levels of manipulation and/or number of

groups, the exhibited sample size issues would be exacerbated further. Another angle that remains to be explored is how perceived role can influence the human-machine interaction in a realistic scenario – one with distractions, greater task difficulty, and significantly longer periods of time in which the task is performed. Since it has been established that we were unable to produce a complacency effect in a short task setting, further research involving enough trials to produce a complacency effect via attention decrement could be used to determine whether perceived role can modulate – as opposed to create – complacency. Furthermore, the findings of a such a study would be of higher external validity, as it would be more representative of applied face matching.

Our results show that perceived role does not have a significant effect on human behaviour in the human-machine interaction, nor does it appear to affect performance, trust in automation, or complacency in short face matching tasks. Whether the lack of an observed complacency effect was due to the task not meeting the criteria for complacency to occur or whether perceived role does not meaningfully affect complacency is unknown. Otherwise, our findings demonstrate that human-AFRS pairs significantly outperform humans alone at unfamiliar face matching tasks. Furthermore, human-AFRS pairs still tend to perform worse than AFRS alone. These findings indicate that despite instructing participants to only interfere with decision when they believed the AFRS erred, they still – on average – curtailed the performance of the AFRS. This effect was present regardless of whether participants were told to use the AFRS as a task aid or perform oversight of the AFRS. This raises the question as to how human operators can be utilised in face matching tasks in a way that does not reduce the overall performance of an AFRS, and whether human operators are necessary at all in face matching scenarios. This problem is exacerbated by the fact that AFRS technology is improving rapidly, such that the gap between human and AFRS performance is increasing over time. If instructing human operators to only overturn AFRS decisions when they believe the AFRS erred still results in them overturning enough correct decisions to curtail its performance, then achieving a human-AFRS team that consistently outperforms AFRS alone seems increasingly difficult.

References

Alenezi, H. M., & Bindemann, M. (2013). The effect of feedback on face-matching accuracy. *Applied Cognitive Psychology*, *27*(6), 735-753.

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: Enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ*, *3*, e1184.

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), 4.

Billings, C. E., Lauber, J. K., Funkhouser, H., Lyman, E. G., & Huff, E. M. (1976). *NASA aviation safety reporting system* (No. NASA-TM-X-3445).

Bobak, A. K., Hancock, P. J., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology*, *30*(1), 81-91.

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2019). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly journal of experimental psychology*, *72*(4), 872-881.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305-327.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior research methods*, *42*(1), 286-291.

Carragher, D. J., & Hancock, P. J. (2022). Simulated automated facial recognition systems as decision-aids in forensic face matching tasks. *Journal of Experimental Psychology: General*.

Carragher, D. J., Sturman, D., Hancock, P., & Carragher, D. J. (2023). Trust in Automation Influences the Accuracy of Human-Algorithm Teams Performing One-to-One Face Matching Tasks. *Retrieved from psyarxiv. com/hs2nb*.

Carroll, J. M. (1997). Human-computer interaction: psychology as a science of design. *Annual review of psychology*, *48*(1), 61-83.

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. Human factors, *65*(1), 137-165.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human factors*, *44*(1), 79-94.

Fysh, M. C., & Bindemann, M. (2017). Forensic face matching: A review. *Face processing: Systems, disorders and cultural differences*, *4*(6), 1-20.

Fysh, M. C., & Bindemann, M. (2018). Human–computer interaction in face matching. *Cognitive science*, *42*(5), 1714-1732.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, *53*(5), 517-527.

Hart, S. G. (1986). NASA task load index (TLX).

Howard, J. J., 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).

Johannsen, G. (2009). Human-machine interaction. *Control Systems, Robotics and Automation*, *21*, 132-62.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *11*(3), 211-222.

Kim, S., McClean, E. J., Doyle, S. P., Podsakoff, N. P., Lin, E., & Woodruff, T. (2022). The positive and negative effects of social status on ratings of voice behavior: A test of opposing structural and psychological pathways. *Journal of Applied Psychology*, *107*(6), 951.

Lamont, A. C., Stewart-Williams, S., & Podd, J. (2005). Face recognition and aging: Effects of target age and memory load. *Memory & cognition*, *33*, 1017-1024.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), 153-184.

Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, *20*(4), 440-458.

Madhavan, P., 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.

Mayer, B., Fuchs, F., & Lingnau, V. (2023). Decision-Making in the Era of AI Support-How Decision Environment and Individual Decision Preferences Affect Advice-Taking in Forecasts. *Journal of Neuroscience, Psychology, & Economics*, *16*, 1-11.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, *38*(2), 311-322.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of experimental psychology: Applied*, *6*(1), 44.

Moreton, R. (2021). Forensic Face Matching. *Forensic Face Matching: Research and Practice; Oxford Academic: Oxford, UK*, 144.

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

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recogniser?. In *Face processing: Systems, disorders and cultural differences* (pp. 173-201). Nova Science Publishers Inc.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors*, *52*(3), 381-410.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. Journal of cognitive engineering and decision making, *2*(2), 140-160.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ... & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171-6176.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PloS one*, *11*(2), e0150036.

Ross, J. M., Szalma, J. L., Hancock, P. A., Barnett, J. S., & Taylor, G. (2008, September). The effect of automation reliability on user automation trust and reliance in a search-and-rescue scenario. In *proceedings of the human factors and ergonomics society annual meeting* (Vol. 52, No. 19, pp. 1340-1344). Sage CA: Los Angeles, CA: Sage Publications.

Schyns, B., Lagowska, U., & Braun, S. (2022). Me, me, me: Narcissism and motivation to lead. *Zeitschrift für Psychologie*, *230*(4), 330.

Stacchi, L., Huguenin-Elie, E., Caldara, R., & Ramon, M. (2020). Normative data for two challenging tests of face matching under ecological conditions. *Cognitive research: principles and implications*, *5*, 1-17.

Thompson, P. S., & Bolino, M. C. (2018). Negative beliefs about accepting coworker help: Implications for employee attitudes, job performance, and reputation. *Journal of Applied Psychology*, *103*(8), 842.

Towler, A., Kemp, R. I., & White, D. (2017). Unfamiliar face matching systems in applied settings. *Face processing: systems, disorders and cultural differences. New York: Nova Science Publishing, Inc*.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS one*, *9*(8), e103510.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1814), 20151292.

White, D., Guilbert, D., Varela, V. P., Jenkins, R., & Burton, A. M. (2022). GFMT2: A psychometric measure of face matching ability. *Behavior research methods*, *54*(1), 252-260.

Xiao, Y. M., Wang, Z. M., Wang, M. Z., & Lan, Y. J. (2005). The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index. *Zhonghua lao dong wei sheng zhi ye bing za zhi= Zhonghua laodong weisheng zhiyebing zazhi= Chinese journal of industrial hygiene and occupational diseases*, *23*(3), 178-181.