The Effects of Making Independent Judgements and Time Pressure on Human Use of Automated Facial Recognition Systems

[redacted]

School of Psychology, University of Adelaide

September 2023

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

Word count: 6,938

Table of Contents

List of Figures

List of Tables

Abstract

Automated Facial Recognition Systems (AFRS) are commonly used in airports to help verify the identity of individuals. The use of AFRS in airports is often teamed with human observers who assist in validating inconclusive AFRS decisions. This verification usually occurs under considerable time pressures. Although previous research has explored time pressure effects on human accuracy in face matching tasks, to date, there has been no research conducted on how time pressure influences human reliance on AFRS. Therefore, this research aimed to investigate how humans use AFRS, and whether time pressure affects this relationship. We first validated a new trial procedure which measured reliance on AFRS through tracking identification decision change. For each pair of stimuli, participants first made an independent identification decision. An AFRS decision was then presented on screen, followed by participants submitting a final response. We applied our new trial procedure to a time pressured task. Each participant (*n* = 56) completed the face matching task in trial blocks where stimuli were presented for 2 seconds, 5 seconds, and 10 seconds. Participant accuracy significantly improved when guided by AFRS compared to independent human decisions. Time pressure produced a non-significant effect on human performance, although, we observed significantly lower mismatch trial accuracy in higher time pressures than in lower time pressures. Furthermore, across both experiments, participants often relied on AFRS, even when the AFRS displayed an incorrect identification decision. Our results have implications for the role of humans performing oversight of AFRS, and potentially limiting the performance of automation.

Keywords: face matching, AFRS assistance, time pressure, reliance on AFRS.

Declaration

This thesis contains no material which has been accepted for the award of any other degree of 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	
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	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 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	

The data in Experiment 1 were collected using discretionary start-up funds provided to my supervisor by the University of Adelaide.

The Effects of Making Independent Judgements and Time Pressure on Human Use of Automated Facial Recognition Systems

Considering the large quantities of people who cross international borders daily, accurately classifying identity is vital to the security of a country (Department of Home Affairs, 2023). In border control, faces are used to verify the identity of individuals through passports. However, international airports are subject to many security and criminal threats through the misuse of identification documents (International Civil Aviation Organisation, 2009). The most common exploit within border control is impersonation based on genuine passports (National Crime Agency, 2015). These imposters may steal a real document based on their similar appearance and attempt to cross international borders on false pretences. Thus, in attempting to mitigate the issue of imposters, forensic face matching occurs. This process relies on the observer determining whether an image is of the same person presenting the passport, to conclude an identity 'match' or 'mismatch'.

Given the importance of face matching in applied contexts, matching unfamiliar faces has been studied extensively throughout the past three decades (e.g., Kemp et al., 1997; O'Toole et al., 2007; Fysh & Bindemann, 2018). Though it might seem easy, matching unfamiliar faces is quite a complex task (Fysh & Bindemann, 2018; Johnston & Bindemann, 2013). Alarmingly, the face matching ability of professionals with extensive experience in validating identification can be similar to individuals who have never completed a face matching task (Weatherford et al., 2021), such that they miss a high ratio of fraudulent passports (Wirth & Carbon, 2017; White et al., 2014). Face matching ability is impacted by illumination (Hill & Bruce; Burton et al., 2005; Jenkins & Burton, 2011), view point of the faces (Bindemann et al., 2013; Estudillo & Bindemann, 2014), image quality (Strathie & McNeill, 2016), individual differences in face memory (Fysh, 2018), time passage between images (Megreya et al., 2013), and information presented on the identification document (Trinh et al., 2022;

Feng & Burton, 2021). Even in controlled and favourable conditions, humans make up to 20-30% of errors (Kemp et al., 1997; Burton et al., 2010). These high human error rates and increased security threats contributed to the development of computer-based face matching algorithms (ICAO, 2009).

Computer-based algorithms, such as Automated Facial Recognition Systems (AFRS), compare features between two images to produce a similarity score, which can be used to determine whether the images are a match or mismatch (Mann & Smith, 2017; Noyes & Hill, 2021). Computers are not subject to the same constraints as humans, such as deterioration of performance through time spent on the task (Alenezi & Bindemann, 2013; Alenezi et al., 2015; Fysh & Bindemann, 2017) or time pressures (Bindemann et al., 2016; Fysh & Bindemann, 2017). Thus, AFRS are useful in scenarios of mass screening, and have now been installed in many major airports. At present, AFRS accuracy can perform better than most humans (O'Toole et al., 2007; Carragher & Hancock, 2022), and similar to expert face matchers (Phillips & O'Toole, 2014; White et al., 2015). Currently, AFRS is generally used as a first point of contact for passport control, where AFRS will display a decision for a human to oversee (Noyes & Hill, 2021). This process is often referred to as human-computer teaming and has attracted research into the dynamics of this interaction (e.g., O'Toole et al., 2007; Fysh & Bindemann, 2018; Howard et al., 2020; Barragan et al., 2022; Carragher & Hancock, 2022).

Human-Computer Teaming in Face Matching

Although face matching has been studied extensively, human-computer interaction in face matching tasks has received little attention. One paper which explored human ability to match faces concluded that face matching decisions are biased by onscreen identification labels (Fysh & Bindemann, 2018). These results were further supported by Howard et al. (2020) who discovered that labels presented by computers and humans both influenced participant judgements. In both studies, participants were required to view a pair of faces which had an identification label present on the screen and decide whether the pair were a match or mismatch. Fysh and Bindemann (2018) set AFRS accuracy at 60%, while 20% of trials were labelled incorrectly, and 20% were labelled ‘unresolved’. Results indicated higher accuracy in trials that were correctly labelled by AFRS than in trials which were incorrectly labelled, suggesting a bias to follow AFRS decisions. These results were consistent when Fysh and Bindemann (2018) instructed participants to simply review the accuracy of the labels; however, participant accuracy was lower than AFRS, suggesting reluctance to accept every decision. However, the low accuracy of the AFRS was not representative of applied settings and could have impacted participant use of automation. Furthermore, Fysh and Bindemann (2018) did not obtain any independent judgements, therefore not demonstrating accuracy of the human without AFRS assistance.

Consequently, Carragher and Hancock (2022) ran a face matching study using a real AFRS and manipulated one match and mismatch trial to display an incorrect identity decision (95% accuracy). Carragher and Hancock (2022) discovered that, while AFRS assistance improved human accuracy, it was still significantly lower than that of AFRS. Similarly, a study exploring the effects of face masks on human-AFRS reliance displayed AFRS accuracy of 95% (Barragan et al., 2022). Participants responded on a 7-point confidence scale. Results indicated higher shifts in confidence scores toward the AFRS when face masks were present, than when masks were absent (Barragan et al., 2022). Considering face masks arguably make the task more difficult (Carragher & Hancock, 2020), it begs the question of whether other factors which increases task difficulty would also increase reliance on AFRS. One such factor may be time pressure; yet, to date there have been no studies exploring the effect time pressure has on human reliance on AFRS.

Time Pressure

Reliance on AFRS in time pressured conditions may be particularly relevant for passport officers' task of validating inconclusive AFRS decisions in high time pressures. Although no research has explored how time pressure affects human reliance on AFRS, research has shown that time pressure often decreases performance in face matching tasks (Ozbek & Bindemann, 2011; White et al., 2015; Bindemann et al., 2016; Fysh & Bindemann, 2017; Wirth & Carbon, 2017). For example, Ozbek and Bindemann (2011) discovered near chance performance in trials where stimuli were displayed for 0.2s, compared to near 90% accuracy when displayed for 2s or no time limit. Furthermore, White et al. (2015) discovered deteriorating performance when stimuli were displayed for 2s compared to 30s. However, these studies only measured very high time pressures of 2 seconds and less, or compared performance between a large range of viewing times.

Subsequent research explored time pressure effects of 2s, 4s, 6s, 8s, and 10s (Bindemann et al., 2016; Fysh & Bindemann, 2017). Time pressure was administered via a speed bar, indicating whether participants were responding within the time of a specific block (i.e., 2s), or responding too fast or slow. Therefore, time pressure was suggested rather than enforced. Accuracy significantly deteriorated in the 2s trial block when compared with performance in the 8s and 10s trial blocks (Bindemann et al., 2016). Similarly, Fysh and Bindemann (2017) reported lower accuracy in 2s and 4s trial blocks when compared with performance in 10s trial blocks. Bindemann et al. (2016) also reported significant accuracy improvement when time pressure decreased in mismatch trials, but not for match trials. However, these studies did not strictly limit presentation time through removing stimuli off screen, thus perhaps not encapsulating the full effect of time pressure. Furthermore, time pressures either systematically increased from 10s to 2s or decreased from 2s to 10s. These studies may be more applicable to applied settings if participants experienced randomised time pressure.

Research Aims

There is a current gap in literature exploring human use of AFRS as a decision aid when participants are subject to time demands. Therefore, this research seeks to investigate the effects of time pressure on human use of AFRS in timed face matching tasks.

Experiment 1

Prior to examining time pressure effects on human use of AFRS, we aim to improve the method used in AFRS assisted face matching tasks. Carragher and Hancock (2022) measured unaided performance through trials where participants selected an identification decision without AFRS assistance. However, Carragher and Hancock (2022) also measured aided performance where participants selected an identification decision while viewing the AFRS response. Thus, Carragher and Hancock (2022) were able to explore how AFRS affected human accuracy by comparing unaided performance with aided performance. However, they were unable to directly measure reliance on AFRS as unaided performance was obtained with different stimuli to that of the aided trials. Without obtaining an independent response before displaying the AFRS response, Carragher and Hancock (2022) could not measure the influence of AFRS on human decisions. Consequently, research using this old trial procedure was unable to explore reliance on AFRS through measuring conformity to an AFRS decision.

Therefore, we sought to validate a new trial procedure in which we will collect two participant decisions for each trial. First, participants will give an initial “same” or “different” response (unaided). After submitting a response, participants will view the AFRS decision and submit a final response of the team (aided). This procedure allows us to measure the frequency of decision change, and whether this change relied on, or rejected, the AFRS decision. We sought to determine whether altering this trial procedure will result in changes to human face matching performance, or whether this procedure could be used in further research.

The potential limitation to our new trial procedure stems from prior research suggesting reluctance to conform to AFRS labels when tasked with reviewing the accuracy of its decision (Fysh & Bindemann, 2018). Thus, we suspected that participants who submit an initial decision prior to viewing one of AFRS would be reluctant to change their decision (i.e., less reliant). Considering the AFRS is highly accurate, lower reliance on its decisions could lead to lower accuracy improvement from initial (unaided) to final (aided) decisions. As such, we hypothesised that there would be a significant interaction between trial procedure and decision type, such that there will be greater accuracy improvement from unaided to aided decisions among participants in the old trial procedure than participants in the new procedure. Conversely, lower reliance on the system may also lead to increased ability to overrule AFRS decisions that are incorrect. Thus, we hypothesised that participants in the new trial procedure would show a higher frequency of overruling incorrect AFRS decisions compared to participants in the old trial procedure.

Method

Participants

Ethical approval was gained from the Human Research Ethics Sub-Committee in the School of Psychology (H-2019-23/01). All participants were recruited through an online research recruitment system called *Prolific* (www.prolific.co.uk). We utilized convenience sampling, as the population were self-selected and compensated with a small payment. A priori power analysis indicated that 69 participants would be required in each trial procedure condition to achieve 80% power and detect an interaction of $\eta^2 = .20$ with an alpha set at .05 (RStudio Team, 2020). Thus, we recruited 81 participants for the new trial procedure condition. However, data were excluded from 2 participants who completed the task in more than 50 minutes, 2 participants who failed an attention check (see below), 2 participants who did not complete the task, and 1 participant who accessed the experiment twice. Our final sample consisted of 74 participants in the new trial procedure condition. Participants in the old trial procedure condition were sampled through a previous face matching study by Carragher, Sturman, and Hancock (In Prep). The data of 74 participants in the old trial procedure condition were randomly selected to match the sample size.

The total sample population included 148 USA residents who were fluent in English. In the new trial procedure condition, participants (38 female, 34 male, 2 other gender) were aged between 20 and 74 (*M* = 38.6, *SD* = 13.26). In the old procedure condition, participants (36 female, 36 male, 2 other gender) were aged between 19 and 72 (*M* = 37.66, *SD* = 12.15).

Design

Utilising an experimental mixed-measures design, we conducted a face matching experiment exploring the effects of a new trial procedure in the human-AFRS teaming paradigm. A between-subjects two-level factor of trial procedure compared accuracy between the new and old trial procedures. A within-subjects two-level factor of decision type compared accuracy between initial (unaided) decisions and final (aided) decisions.

Materials

The Glasgow Face Matching Task (GFMT2-S)

The Glasgow Face Matching Task – Short (GFMT2-S; White et al., 2022) was used to assess face matching ability. For each of the 80 trials in the GFMT2-S, two images with a face in each were displayed on screen. Participants must decide whether the two images show the same person (match trial), or two different people (mismatch trial). Responses on each trial were made by selecting a “same” or “different” response. Trial pairs consisted of either a male pair or female pair and varied in image quality. The 80 trials were evenly split into Trial Block A (*n* = 40) and Trial Block B (*n* = 40) of equal difficulty, with 20 match and 20 mismatch trials in each. Trial block presentation was counterbalanced across participants to minimise any effects from time spent on the task. To further minimise the effects of time spent on the task, trial presentation was randomised to ensure each participant viewed trials in a random order.

Face Matching Trial Procedure

"Old" Procedure. Carragher et al. (In Prep) utilised a trial procedure in which participants were presented trial pairs in the GFMT2-S with a predetermined AFRS decision. These AFRS-assisted trials will be measured as 'final' trials. To measure baseline performance, participants completed separate trials without the assistance of AFRS. Baseline performance will be measured as 'initial' trials.

"New" Procedure. In our new experiment, we altered the trial procedure by requiring participants to submit an initial judgement prior to viewing the predetermined AFRS decision. Participants then viewed the AFRS decision on the same trial and submitted a final response.

Automated Facial Recognition System

The simulated AFRS was based on the performance of the real Deep Convolutional Neural Network (DCNN) that was used in Carragher and Hancock (2022). The accuracy of the real DCNN on the entire GFMT2-S was 97.5%, correctly identifying 78/80 trials. Additional AFRS errors were added by manipulating two trials to state an incorrect identity decision. The match and mismatch trial in which the DCNN produced the lowest similarity score were chosen. This gave participants an opportunity to overrule incorrect decisions made by the AFRS. This manipulation resulted in AFRS accuracy of 95% in both trial procedures.

Attention Checks

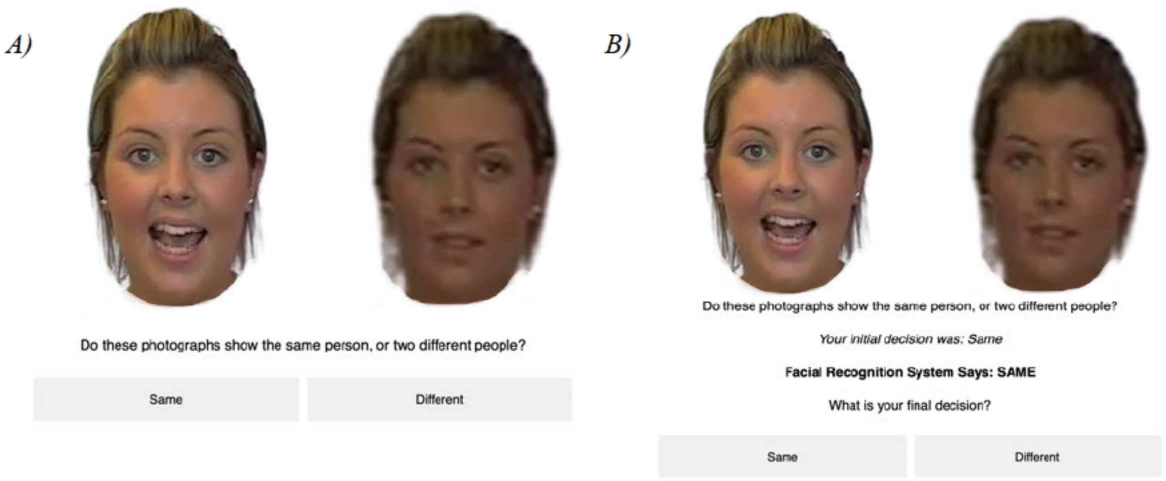
Attention checks occurred in each block to allow screening for inattentive participants. The attention check displayed two images of easily recognisable celebrities who have distinct characteristic differences. Given the easily recognisable differences between presented celebrities, participants failed the attention check if they did not select "different" as their response. The attention checks were the same in both experiments, with an additional attention check present in the third trial block of our second experiment.

Procedure

This study took place online using Qualtrics survey software to run the experiment. After providing informed consent, participants filled in a short demographic questionnaire. Embedded in the task instructions was the AFRS accuracy (95%). Two images from the GFMT2-S were presented on screen simultaneously, with the question, "*Do these photographs show the same person, or two different people?*" (see Figure 1). The two images remained on screen until participants selected either "same" or "different". Participants were then shown their initial response and the decision of AFRS and were asked to submit a final decision. Participants were given the opportunity to take a self-paced break halfway through the experiment. In line with ethics, debriefing occurred at conclusion of the task.

Figure 1

Example Match Trial of GFMT2-S During Initial (A) and Final (B) Decisions



Do these photographs show the same person, or two different people?

Do these photographs show the same person, or two different people?

Your initial decision was: Same

Facial Recognition System Says: SAME

What is your final decision?

Note. Stimuli remained on screen until a response was given.

Analysis

Accuracy

Accuracy performance was measured by creating an average accuracy percentage for each participant's initial and final decisions in both trial procedures [((Correct Identification Decision / Total Trials)*100)].

Decision Change, Reliance, and Rejection

Given the new trial procedure included an initial and final decision for each trial pair, we were able to create a measure of decision change between an initial and final response. On each trial, decision change occurred when the participant's final identification decision was different to their initial identification decision. Reliance occurred when the participant changed their decision to match the AFRS. Rejection occurred when the participant changed their decision against AFRS. Decision change, reliance, and rejection were measured through creating an average percentage of frequency [e.g., ((Decision Change / Total Trials)*100)]. Given these measures were only available for the new trial procedure, they were explored through descriptive statistics rather than formal statistical analyses. JASP (2023) was used for all statistical analyses.

Results

Assumption Checks

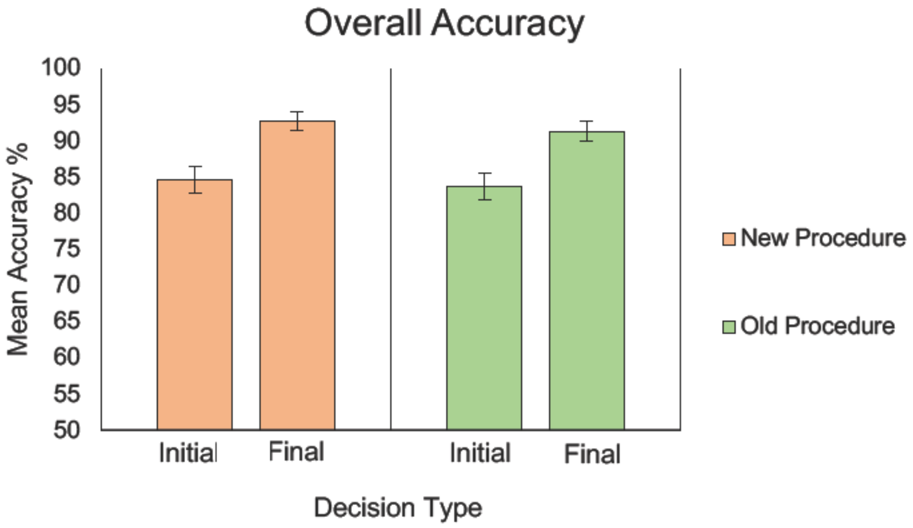
Prior to analysis, data were tested for normality to ensure assumptions of a parametric ANOVA were upheld. The data failed a Shapiro-Wilk test of normality, indicating abnormal distribution. However, given the sensitivity of the Shapiro-Wilk test to large sample sizes ($n > 50$) (Mishra et al., 2019), these results were interpreted with caution. Further visualisation of the data showed slight negative skews for all factors, which indicated clustering of scores at the high end of accuracy. Given the robust nature of ANOVA to violations of normality (Blanca et al., 2017), it is likely that skewness will not make a fundamental difference in our analyses (Tabachnick et al., 2013). Greenhouse-Geisser sphericity corrections were applied when necessary. Therefore, we ran our statistical analyses as planned.

Trial Procedure

To test our primary hypothesis that trial procedure and AFRS assistance impact performance, we performed a 2 x 2 mixed ANOVA on mean accuracy scores between trial presentation and decision type. A significant main effect of decision type was obtained ($F(1,146) = 128.88$, $p < .001$, $\eta^2 = .203$). Initial decisions obtained mean accuracy of 84.13% (*SD* = 9.18), whereas final decisions obtained mean accuracy of 92.06% (*SD* = 6.32) (see Figure 2). The main effect was non-significant for trial procedure ($F(1,146) = 1.18$, $p = .278$, $\eta^2 = .005$), and the interaction was non-significant between trial presentation and decision type ($F(1,146) = 0.10$, $p = .745$, $\eta^2 <.001$). These results show that trial procedure did not affect aided performance on the task. However, whether the participant was aided by AFRS impacted performance.

Figure 2

Overall Face Matching Accuracy Across Trial Procedure Conditions



Note. Error bars represent 95% confidence intervals for standard error of the mean.

Incorrect AFRS Decisions

To test our secondary hypothesis that trial procedure would impact frequency of overruling incorrect AFRS decisions, we assessed mean accuracy on trials where the AFRS made an incorrect identification decision. In the new trial procedure, participants obtained an average accuracy of 64.53% in trials where the AFRS displayed an incorrect identification decision. In the old trial procedure, participants obtained an average accuracy of 68.24% in trials where the AFRS displayed an incorrect identification decision. These results imply participants in the old trial procedure may have had a higher frequency of overruling incorrect AFRS decisions, contrary to our hypothesis.

Exploratory Reliance and Rejection Analysis

Lastly, we ran exploratory analyses on data collected from the new trial procedure to examine descriptive statistics on decision change, reliance on the AFRS, and rejection of the AFRS. In the new trial procedure, participants changed their final identification decision from their initial on 10.49% of trials. In trials where decision change occurred, participants relied on AFRS (10.02%) more than they rejected AFRS (0.47%). On trials where the AFRS displayed a correct decision, decision change (10.28%), reliance (9.79%), and rejection (0.48%) occurred less than during trials where the AFRS displayed an incorrect decision. During incorrect AFRS decision trials, identification decision change occurred on 14.53% of trials. Of these trials, participants relied on the AFRS (14.19%) more than they rejected AFRS (0.34%). These results show that reliance on AFRS is more likely to occur than rejection, regardless of whether the AFRS displays a correct decision.

Discussion

Contrary to our predictions, the results demonstrate that the new trial procedure does not impact aided face matching performance. This statement is supported through participants obtaining similar mean accuracy regardless of trial procedure, and similar ability to overrule incorrect AFRS decisions. However, across both trial procedures, we saw participants improve their accuracy when assisted by AFRS. While both of our hypotheses were not supported, this does mean that we are able to incorporate the new trial procedure in future face matching experiments. The new trial procedure will allow opportunity to measure decision change, reliance, and rejection of AFRS. We will now use the new trial procedure to investigate the effect of time pressure on human use and reliance on AFRS.

Experiment 2

After validating the new trial procedure, we applied it to a time pressured face matching task. Border Control Officers work under considerable time pressures in airports (Department of Home Affairs, 2022; Australian National Audit Office, 2010). Time pressure has been shown to reduce face matching accuracy in controlled laboratory environments (Ozbek & Bindemann, 2011; Bindemann et al., 2016; Wirth & Carbon, 2017; Fysh & Bindemann, 2017). Additionally, prior research has shown that humans are more likely to conform to automated decisions when a task is more difficult (Weger et al., 2015). Therefore, we aimed to investigate how time pressure would affect human reliance on decisions from an AFRS.

Human face matching accuracy deteriorates when stimuli are viewed for two seconds or less (Fysh & Bindemann, 2017). Additionally, several studies have reported face matching to occur, on average, within 6 seconds (e.g., Estudillo & Bindemann, 2014; Papesh & Goldinger, 2014). Therefore, we expect to see lower initial accuracy when trials are presented for 2 seconds, compared to when presented for 5 and 10 seconds. Due to the increased difficulty, we expect participants will show greater reliance on the highly accurate AFRS (92.3%). As such, we hypothesise that there will be a significant interaction between time pressure and decision type, such that the 2 second trial block will have a higher magnitude of accuracy improvement from initial to final decisions than the 5 second or 10 second trial blocks. Finally, we hypothesise that there will be a significant interaction between confidence and decision change in each time pressure condition, where participants will be more likely to change their response to rely on AFRS when they are less confident in their response[1].

Method

Participants

Participants were recruited through the online Research Participation System (RPS) at the University of Adelaide. A priori power analysis indicated 52 participants would be required to achieve an 80% power to detect an interaction of $\eta^2 = .25$ with an alpha set at .05 (RStudio Team, 2020). The sample consisted of 56 first year undergraduate psychology students (46 female, 9 male, 1 other gender) who participated in exchange for course credit. Participants were aged between 17 and 45 years old (*M* = 20, *SD* = 4.2) and were fluent in English. Participants were ineligible to complete the study if they had participated in another face matching experiment during the semester. Exclusion criteria remained the same as our first experiment.

Design

Utilising an experimental mixed-subjects design, we used the new response method and applied it to a task where participants were subject to a high pressure (2 second), average pressure (5 second), and low pressure (10 second) time condition. A three-level factor of time pressure compared accuracy between high, average, and low pressure trial blocks. A two-level factor of decision type compared accuracy between initial and final decisions. Another two-level factor of match type compared accuracy in match and mismatch trials. Finally, a two-level factor of decision change compared frequency of occurrence (no change and change).

[1] We originally hypothesised a significant interaction between confidence and decision type in each time pressure block. The factor of decision type was changed to the factor of decision change to better suit the investigation of confidence change and reliance on AFRS.

Materials

The Glasgow Face Matching Task (GFMT2-S)

Identical to Experiment 1, participants completed the GFMT2-S. However, two trials were removed to create three even trial blocks among time pressure conditions. The match and mismatch trials with the highest accuracy reported by the GFMT2-S creators (White et al., 2022) were removed, leaving a total of 78 trials (*n* in each group = 26) with 13 match and 13 mismatch trials in each block. Similar to Experiment 1, identification responses were made by selecting a “same” or “different” response. However, our second experiment differed by including a 6-point confidence scale, including “definitely”, “probably”, and “guess”, as demonstrated in Figure 3.

Time Pressure

Presentation of the GFMT2-S occurred in three time pressure conditions. All participants completed a high pressure (2 seconds), average pressure (5 seconds), and low pressure (10 seconds) trial block. Time pressure was enforced by displaying GFMT2-S trials for the allocated time before disappearing off screen.

Automated Facial Recognition System

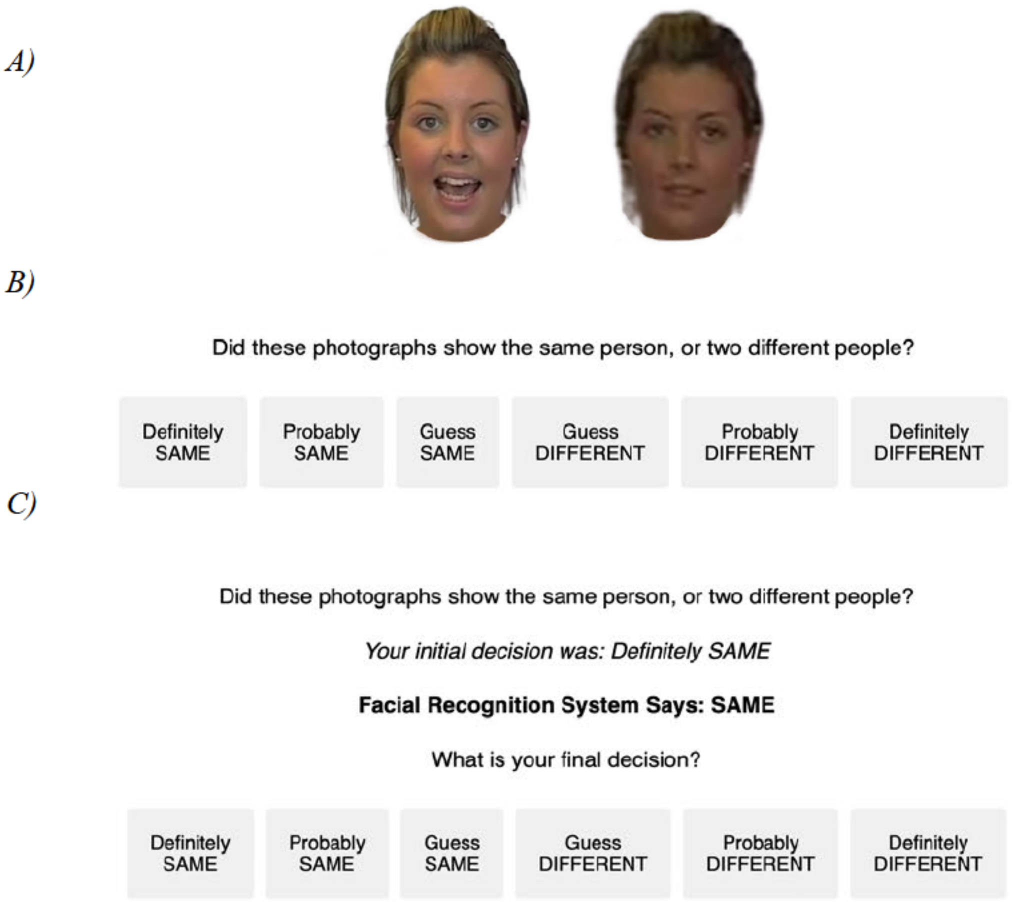
The accuracy of the simulated AFRS was 92.3%, correctly identifying 72/78 trials. The three match and mismatch trials in which the DCNN produced the lowest similarity scores were manipulated to show an incorrect identification decision.

Procedure

Participants were informed the images would be displayed for either 2, 5, or 10 seconds, and they would progress to each trial block without notice. Furthermore, participants were informed that AFRS accuracy was 92.3%. Presentation of the trials differed slightly from our first experiment in that images were only displayed for the time allocated in the specific trial block before disappearing off screen. Once the images had disappeared, participants could provide an initial response to the same prompt question as Experiment 1, however, this was on the confidence scale. Without viewing the images again, participants were shown their initial response and the decision of AFRS before submitting a final response (see Figure 3).

Figure 3

Example Match Trial of GFMT2-S with 6-Point Confidence Identification Decision Scale



Did these photographs show the same person, or two different people?

Did these photographs show the same person, or two different people?

Your initial decision was: Definitely SAME

Facial Recognition System Says: SAME

What is your final decision?

Note. Trial pairs are presented on screen without response scale (A). Example of initial (B) and final (C) decision on the confidence identification scale.

Analysis

Accuracy

Identification responses between 1-3 were recorded as ‘same’, and values 4-6 were recorded as ‘different’. Average accuracy performance for each participant was then calculated the same as our first experiment.

Confidence

Confidence was measured through participants selecting their identification response on a 6-point ‘same’ and ‘different’ scale. Where participants selected a ‘definitely’ response, they were most confident, ‘probably’ was less confident, and ‘guess’ was least confident.

Decision Change

There were two measures of change. One measure was identification decision change, which followed the same procedure as our first experiment. The second was confidence change. Participants were coded as either not having changed their final confidence from their initial confidence category, toward AFRS, or against AFRS.

Results

Assumption Checks

Results of Shapiro-Wilk tests of normality were significant for all variables, except for final accuracy in the low-pressure condition. This indicated that the data did not follow normal distribution. Sphericity corrections were applied where possible, however, we proceed with our analyses on the same basis outlined in Experiment 1.

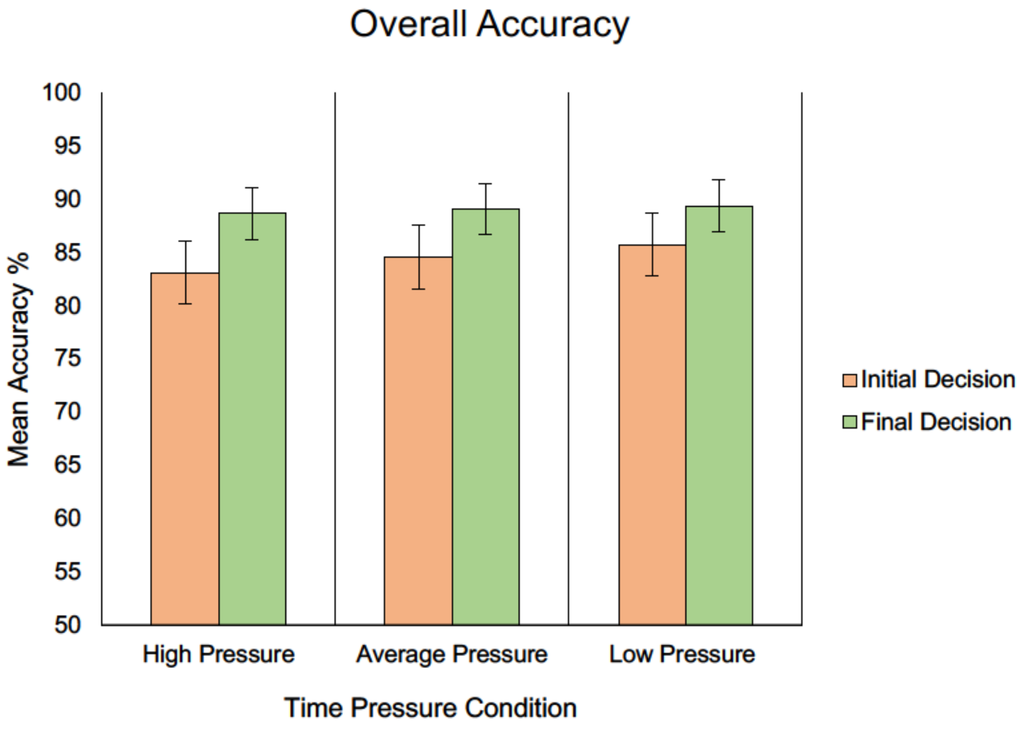
Time Pressure

To test our primary hypothesis that time pressure would impact accuracy performance, we performed a 2 x 3 repeated measures ANOVA on mean accuracy. Factors of the ANOVA included decision type in each time pressure condition. A significant main effect of decision type was obtained ($F(1,55) = 82.91$, $p < .001$, $\eta^2 = .147$). Initial decisions obtained mean accuracy of 84.5% (*SD* = 8.5), whereas final decisions obtained mean accuracy of 89.1% (*SD* = 8.0). The main effect of time pressure was non-significant ($F(2,110)$

= 1.12, $p = .330$, $\eta^2 = .013$), and the interaction between decision type and time pressure was non-significant ($F(2,110) = 2.54$, $p = .083$, $\eta^2 = .005$). These results show that AFRS aid impacted performance on the task, however, time pressure did not (see Figure 4).

Figure 4

Overall Face Matching Accuracy Across Time Pressure Conditions



Note. Error bars represent 95% confidence intervals for standard error of the mean.

Identification Decision Change, Reliance, and Rejection

Exploratory analyses were performed on decision change using a repeated measures ANOVA on frequency of occurrence. On average, participants changed their final identification decision from their initial identification decision on 7.37% of trials. When the AFRS displayed an incorrect correct decision, decision change occurred more (17.56%) than in than in trials where the AFRS displayed a correct decision (6.52%). Participants relied on

AFRS more (7.33%) than they rejected AFRS (0.04%). On the trials where rejection occurred, the AFRS displayed a correct identification decision. Participants appeared to change their decision more in the high pressure condition (8.24%) than the average pressure (6.73%) and low pressure condition (7.14%). Results of the ANOVA returned a non-significant interaction ($F(2,110) = 1.12, p = .328, \eta^2 <.001$), meaning time pressure did not significantly impact frequency of decision change.

Confidence Change, Reliance, and Rejection

Subsequently, to explore our secondary hypothesis that confidence impacts reliance on AFRS in each time pressure condition, a statistical analysis using ANOVA was originally planned. However, not all participants selected each initial confidence category in each time pressure condition, thus resulting in missing data. As such, we chose to examine descriptive statistics of initial confidence category and confidence change. As suspected, participants shifted confidence toward the AFRS the most in trials where they were least confident (see Table 1). Furthermore, participants were more likely to change their confidence, either toward or away from the AFRS, in trials where the AFRS displayed an incorrect identification decision. These results remained relatively consistent across all time pressure conditions (see Figure 5), suggesting that time pressure did not influence confidence change. It is important to note that, when the AFRS was correct, participants who selected “definitely” were able to move toward the AFRS when their initial decision was incorrect. Conversely, when the AFRS was incorrect, participants who selected “definitely” were able to move toward the AFRS when their initial decision was correct.

Table 1

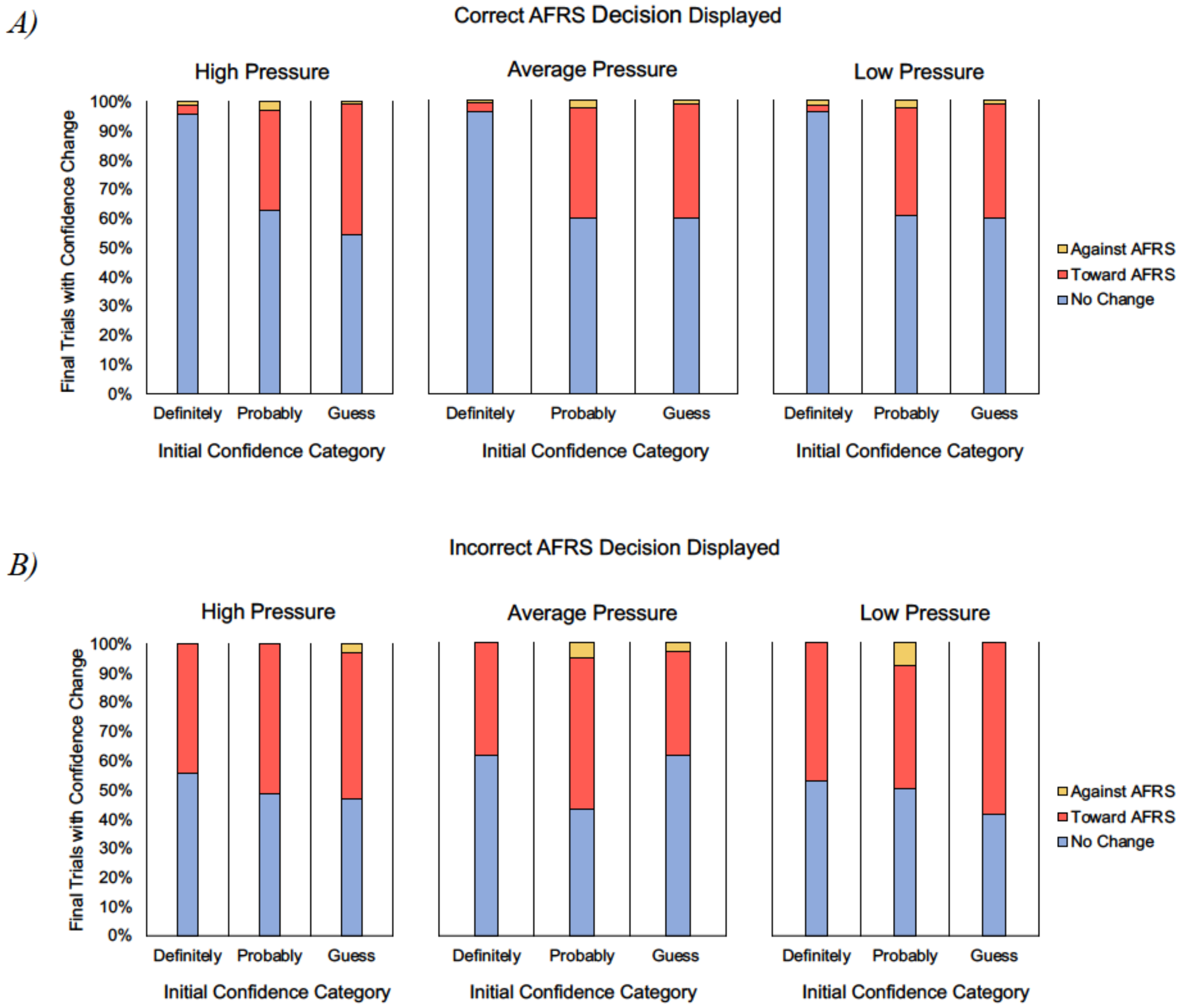
Confidence Change Depending on Initial Confidence

Confidence Change	Initial Confidence Category		
	Definitely	Probably	Guess
	Correct AFRS Decision		
No Change	95.85%	61.13%	57.86%
Toward AFRS	2.84%	36.06%	40.96%
Against AFRS	1.31%	2.82%	1.18%
Total Trials	41.87%	37%	21.13%
	Incorrect AFRS Decision		
No Change	56.8%	47.58%	50.57%
Toward AFRS	43.2%	47.58%	47.13%
Against AFRS	0%	4.84%	2.3%
Total Trials	37.2%	36.9%	25.89%

Note. The numbers displayed in this table demonstrate the percentage of trials within an initial confidence category where confidence change occurred.

Figure 5

Confidence Change in Correct AFRS Trials (A) and Incorrect AFRS Trials (B)



Note. The data in this table represents the percentage of trials in which confidence change occurred per initial confidence category.

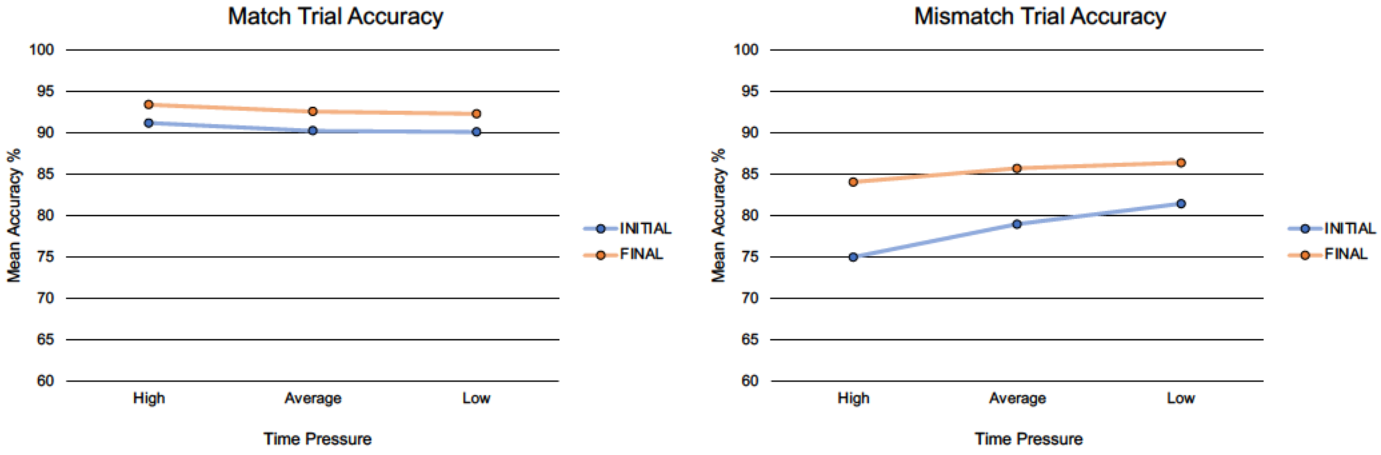
Match Type Accuracy

A 3 x 2 x 2 mixed ANOVA examining differences in mean accuracy of match type and decision type across time pressure conditions was conducted. Significant main effects of match type ($F(1,55) = 45.53, p < .001, \eta^2 = .194$) and decision type ($F(1,55) = 82.93, p < .001, \eta^2 = .043$) were obtained. There was also a significant interaction between match type and decision type ($F(1,55) = 24.91, p < .001, \eta^2 = .011$). The main effect of time pressure was non-significant ($F(2,110) = 1.12, p = .330, \eta^2 = .004$) and the interaction between time pressure, match type, and decision stage was non-significant ($F(2,110) = 2.29, p = .106, \eta^2 = .001$). However, the interaction between time pressure and match type was close to significant ($F(2,110) = 3.05, p = .051, \eta^2 = .011$).

Thus, given the lower than expected sample size potentially influencing the magnitude of significance, we examined differences among match type accuracy in each time pressure condition. Results of a 2 x 2 repeated measures ANOVA examining differences in mean accuracy among match trials obtained a nonsignificant interaction ($F(2,110) = 0.01, p = .989, \eta^2 <.001$). Conversely, a 2 x 2 repeated measures ANOVA examining differences in mean accuracy among mismatch trials obtained a significant interaction ($F(2,110) = 3.588, p = .031, \eta^2 = .007$). These results imply higher time pressure significantly decreased accuracy performance in mismatch trials, but not in match trials (see Figure 6).

Figure 6

Match Type Initial and Final Accuracy



Note. The data points in this figure represent mean accuracy performance in each group.

General Discussion

This study explored the effects of time pressure on human performance and reliance on AFRS in a face matching task. Contrary to our predictions, time pressure did not significantly impact human performance in the task. However, we consistently observed human accuracy improving after viewing an AFRS decision. Furthermore, when participants were less confident in their responses their reliance on AFRS increased. Although, when participants were the most confident in their response and the AFRS contradicted their decision, confidence often dropped to follow the AFRS. Identification decision change occurred in similar frequencies across all time pressures. However, on almost all trials where identification decision change occurred, participants relied on AFRS. Exploratory analyses suggested that time pressure impacted human performance in mismatch trials, but not in match trials. The pattern of these results will be further examined along with the implication to applied settings.

Time pressure was expected to significantly impact overall performance, however, this was not the case. Our results are similar to Ozbek and Bindemann (2011), where accuracy was only impacted by time pressures of less than 1s and remained relatively consistent when viewed for 2s or for an unlimited time. Therefore, our choice of 2s, 5s, and 10s viewing time perhaps influenced the nonsignificant effect. However, our viewing times were chosen as prior research demonstrated lower accuracy in 2s and 4s pressures than in 10s pressures (e.g., Bindemann et al., 2016; Fysh & Bindemann, 2017). In the current study, along with Ozbek and Bindemann (2011), strict time pressure was enforced through stimuli being removed from view. This method differed from Bindemann et al. (2016) and Fysh and Bindemann (2017) who simply suggested time pressure rather than enforcing it. We suggest that perhaps the nature of the time pressure influenced the magnitude of effect. One study determined that global time pressures, where participants must respond to numerous trials within a specified time, i.e., 10 minutes, are more detrimental to performance than strict time pressures (Wirth & Carbon, 2017). It is possible that participants in Bindemann et al. (2016) and Fysh and Bindemann (2017) processed the task similarly to global time pressure, thus resulting in a larger impact.

Additionally, a notable difference between previous findings and the current research is the use of AFRS assistance in final decisions. Despite being non-significant, there was a pattern of gradual increase in initial accuracy as time pressure decreased. These initial decisions are presented similarly to other face matching studies without AFRS assistance (e.g., Bindemann et al., 2016; Fysh & Bindemann, 2017; White et al., 2015) who also reported increased accuracy as time pressure decreased. Interestingly, while initial accuracy appeared to change, final accuracy remained consistent throughout time pressures. These results are fascinating, as initial performance appears to be influenced by time pressure as demonstrated in prior research; whereas final decisions aided by AFRS appear to be unaffected. We suggest that perhaps time pressure influenced reliance on AFRS rather than performance in final trials.

Participant performance significantly improved after viewing an AFRS decision across all time pressures, which suggested reliance on AFRS. While non-significant, it appears participants changed their decision more in the high time pressure. However, participants were more likely to rely on the AFRS during incorrectly labelled trials, thus often failing to overrule the AFRS. These findings are consistent with previous literature where participants often failed to correct AFRS errors (e.g., Carragher & Hancock, 2022; Fysh & Bindemann, 2018; Barragan et al., 2022). Our results are also similar to a study which reported participants changing their initial decision to match an automated response 25% of the time (Salim et al., 2023). However, despite often relying on the AFRS, participants were still unable to achieve the same level of performance as the AFRS. Similarly, Carragher and

Hancock (2022) report that participants were unable to match the accuracy of their simulated AFRS. This lower human accuracy thus implies that participants did not simply follow every AFRS decision; rather, participants relied on AFRS while also overruling correct decisions.

As suspected, participants were more likely to rely on the AFRS when they were less confident in their initial response. Intriguingly, despite being the most confident in their response, participants shifted confidence toward the AFRS when an incorrect AFRS decision was presented. These results are similar to those obtained by Barragan et al. (2022) where participants often shifted confidence toward the system when the task became more difficult. Similarly, Weger et al. (2015) reported increased conformity to automated decisions when the task was more difficult. Thus, we suggest that perhaps participants perceived the task as more difficult after a highly accurate AFRS contradicted their initial response. Human decision making often stems from the pathway requiring least cognitive effort (Parasuraman & Manzey, 2010). Thus, instilling doubt in a confident decision may have encouraged participants to rely on automation. We are now able to measure the magnitude in confidence shifts and reliance on automated decisions, which is a particularly advantageous method while exploring human decision making in the face of technology.

Furthermore, participants struggled more with telling two faces apart with AFRS assistance than determining an identity match without assistance. These results are consistent with prior research reporting a notable difference in mismatch performance compared to match performance (e.g., White et al., 2022; Papesh & Goldinger, 2014; Bindemann et al., 2010). Furthermore, increased time pressure appeared to decrease mismatch trial accuracy but not match accuracy. Similarly, Bindemann et al. (2016) reported significant accuracy decreases in higher time pressures for mismatch trials, but not for match trials. Bindemann et al. (2016) also reported slower response times in these mismatch trials. We therefore suggest that mismatch trials may be perceived as more difficult than match trials. Fysh & Bindemann

(2023) suggest that mismatch trials take more cognitive effort to evaluate than match trials. Given participants were subject to time pressure demands, it is possible that participants were unable to process the necessary information to conclude an identity mismatch, thus perhaps creating a bias to select a match response.

Implications

In applied settings, the performance obtained in our study would result in serious security breaches, with overall human performance averaging less than 90%, even with AFRS assistance. Furthermore, participants failed to reach the same accuracy as the AFRS alone and often failed to correct errors. Thus, our results have implications for the role of humans in forensic face matching, potentially limiting the accuracy of an AFRS and failing to provide accurate oversight of AFRS errors. The effect of time pressure in mismatch trials further emphasizes how placing workers under considerable time constraints may result in more undetected imposters. It is important to consider how increasing time pressure targets of processing in airports may have a detrimental effect on the security of a country. For example, some airports aim to reduce passport processing times of incoming and outbound passengers each year (Department of Home Affairs, 2023).

In our experiments, AFRS decisions had the power to not only influence confidence in a participant decision, but also change a participant's final identification decision. If technology can influence those who are the most confident in their responses and sway a correct human decision to be incorrect, perhaps the role of humans performing oversight of automated decisions should be carefully re-evaluated. Therefore, the demonstrated reliance on AFRS has implications for the way in which humans use technology; and if incorrectly used, we could see more unintended consequences than benefits (Lyon, 2003).

Limitations and Future Directions

A limitation to acknowledge for future research may stem from the high initial accuracy of participants, suggesting the task may have been easier than predicted. White et al. (2022) reported normed accuracy of 75.9% in the validation of the GFMT2-S, with 77% for match trials and 74.9% for mismatch trials. Comparatively, participants in our study achieved accuracy of 85.4%, with 90.5% for match trials and 78.4% for mismatch trials. The possibility of decision change and reliance were reduced when participants were gaining higher than expected initial accuracy. Thus, with AFRS accuracy set at 92.3%, our reliance data may have obtained a smaller effect than if initial accuracy was in greater alignment to the data reported by White et al. (2022). Nevertheless, our results provide valuable insight into the human tendency to rely on automated decisions, despite the automation providing incorrect data.

Furthermore, our choice in sampling undergraduate students may have influenced the effects associated with mismatch trial accuracy. A bias toward selecting a match response rather than mismatch has been observed in numerous studies (e.g., Bindemann et al., 2016; Alenezi & Bindemann, 2013). It is possible that a student sample experienced little incentive to complete the task with seriousness, potentially increasing a match bias as such. Future studies could aim to explore the effects of time pressure and reliance on AFRS in a population who benefits from the use of AFRS. We also suggest that applying a measure of Signal Detection Theory may be useful in examining match type biases.

Directions for future research could focus on applying our new trial procedure to a global time pressured experiment to explore how reliance on AFRS is impacted. In a global time pressure, participants are allocated a certain amount of time to complete all trials within the experiment, i.e., 10 minutes. This suggestion stems from prior research demonstrating a global time pressure to be more detrimental than a strict trial time limit (Wirth & Carbon, 2017). Moreover, global time pressures may be more applicable to many applied settings.

Conclusion

Face matching plays an important role in the safety and security of society (Department of Home Affairs, 2023). Thus, technology advancements have seen the development of AFRS in applied settings (Noyes & Hill, 2021) which are often more accurate than humans (Carragher & Hancock, 2022). By validating a new trial procedure, we improved the method used in AFRS assisted face matching tasks. After creating a measure of decision change, reliance, and rejection, we have shown that participants are likely to rely on the identification labels of AFRS, even when the AFRS displays an incorrect result. Concerningly, during incorrectly labelled trials, participant accuracy often decreased from relying on AFRS after initially selecting a correct identification decision. However, participants consistently failed to reach the accuracy of the AFRS through overruling correct decisions and failing to overrule errors. Moreover, time pressure influenced ability to distinguish two different faces. Though, performance in mismatch trials were significantly lower across all conditions. This match type effect has implications for applied settings where the purpose of face matching is often to catch people who are using false identification documents. Consideration must be taken in determining the role of humans in forensic face matching while working with AFRS or performing oversight of AFRS decisions. Perhaps future implementation of automated systems could be accompanied with extensive training on how humans can best use this advancement.

Reference List

Alenezi, H., & Bindemann, M. (2013). The Effect of Feedback on Face-Matching Accuracy. *Applied Cognitive Psychology, 27*(6), 735–753. https://doi.org/10.1002/acp.2968

Alenezi, H., 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, 2015*(8), e1184. https://doi.org/10.7717/peerj.1184

Australian National Audit Office. (2010). *Processing of International Air Passengers*. Australian Customs and Border Protection Service.

Barragan, D., Howard, J. J., Rabbitt, L. R., & Sirotin, Y. B. (2022). COVID-19 masks increase the influence of face recognition algorithm decisions on human decisions in unfamiliar face matching. *PloS one, 17*(11), e0277625. https://doi.org/10.1371/journal.pone.0277625

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The Effect of Image Pixelation on Unfamiliar-Face Matching. *Applied Cognitive Psychology, 27*(6), 707–717. https://doi.org/10.1002/acp.2970

Bindemann, M., Avetisyan, M., & Blackwell, K.-A. (2010). Finding Needles in Haystacks: Identity Mismatch Frequency and Facial Identity Verification. *Journal of Experimental Psychology. Applied, 16*(4), 378–386. https://doi.org/10.1037/a0021893

Bindemann, M., Fysh, M., Cross, K., & Watts, R. (2016). Matching faces against the clock. *i-Perception, 7*(5), 1-18. https://doi.org/10.1177/2041669516672219

Blanca, M., Alarcon, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-Normal Data: Is ANOVA Still a Valid Option? *Psicothema, 29*(4), 552-557. https://doi.org/10.7334/psicothema2016.383

Burton, M., Jenkins, R., Hancock, P., White, D. (2005). Robust representations for face

Recognition: The power of averages. *Cognitive Psychology, 51*(3), 256–284.

Burton, M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behaviour Research Methods, 42*, 286-291.

Carragher, D., & Hancock, P. (2020). Surgical face masks impair human face matching performance for familiar and unfamiliar faces. *Cognition Research 5*(59). https://doi.org/10.1186/s41235-020-00258-x

Carragher, D. & Hancock, P. (2022). Simulated Automated Facial Recognition Systems as Decision-Aids in Forensic Face Matching Tasks. *Journal of Experimental Psychology: General*, *152*(5), 1286-1304. https://doi.org/10.1037/xge0001310

Department of Home Affairs. (2022). *2021-2022 Annual Report*. Australian Government.

Department of Home Affairs. (2023). *2022-2023 Corporate Plan.* Australian Government.

Estudillo, A., & Bindemann, M. (2014). Generalization across View in Face Memory and Face Matching. *i-Perception (London), 5*(7), 589–601. https://doi.org/10.1068/i0669

Feng, X., & Burton, M. (2021). Understanding the document bias in face matching. *Quarterly Journal of Experimental Psychology, 74*(11), 2019–2029. https://doi.org/10.1177/17470218211017902

Fysh. (2018). Individual differences in the detection, matching and memory of faces. *Cognitive Research: Principles and Implications, 3*(1), 1-12. https://doi.org/10.1186/s41235-018-0111-x

Fysh, M., & Bindemann M. (2017). Effects of time pressure and time passage on face-matching accuracy. *Royal Society Open Science. 42*(5), 1714-1732.

Fysh, M., & Bindemann, M. (2018). Human-Computer Interaction in Face Matching. *Cognitive Science, 42*(5), 1714-1732.

Fysh, M., & Bindemann, M. (2023). Understanding Face Matching. *Quarterly Journal of Experimental Psychology, 76*(4), 862-880.

Hill, H. & Bruce, V. (1996). Effects of Lighting on the Perception of Facial Surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 986-1004. doi: 10.1037/0096-1523.22.4.986.

Howard, 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), Article e0237855, 1-18. https:// doi.org/10.1371/journal.pone.0237855

International Civil Aviation Organisation (ICAO). (2009). *MRTD Report: Defending the Document, 4*(2), Global Enterprise Technologies Group.

JASP Team (2023). JASP (Version 0.17. 3) [Computer software].

Jenkins, R., & Burton, M. (2011). Stable face representations: Philosophical Transactions of the Royal Society of London. *Biological Sciences, 366*(1571), 1671–1683. https://doi.org/10.1098/rstb.2010.0379

Johnston, R. & Bindemann, M. (2013). Introduction to Forensic Face Matching. *Applied Cognitive Psychology, 27*(6), 697-699. https://doi.org.10.1002/acp.2963

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.

Lyon, D. (2003). Technology vs 'terrorism': circuits of city surveillance since September 11th. *International Journal of Urban and Regional Research, 27*, 666-678. https://doi.org/10.1111/1468-2427.00473

Mann, & Smith, M. (2017). Automated facial recognition technology: recent developments and approaches to oversight. *University of New South Wales Law Journal, 40*(1), 121–145. https://doi.org/10.53637/KAVV4291

Megreya, A., Sandford, A., & Burton, A. M. (2013). Matching Face Images Taken on the

Same Day or Months Apart: the Limitations of Photo ID. *Applied Cognitive Psychology, 27*(6), 700–706. https://doi.org/10.1002/acp.2965

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia, 22*(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

National Crime Agency (NCA). (2015). *National strategic assessment of serious and organised crime 2015.*

Noyes, E., & Hill, M. Q. (2021). Automatic recognition systems and human computer interaction in face matching. *Forensic face matching: Research and practice,* 193–215. https://doi.org/10.1093/oso/9780198837749.003.0009

O'Toole, A. J., Abdi, H., Jiang, F., & Phillips, P. J. (2007). Fusing face recognition algorithms and humans. *IEEE: Transactions on Systems, Man & Cybernetics, 37*(5), 1149-1155.

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research (Oxford), 51*(19), 2145–2155. https://doi.org/10.1016/j.visres.2011.08.009

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception, & Psychophysics, 76*(5), 1335–1349. https://doi.org/10.3758/s13414-014-0630-6.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32*(1), 74–85. https://doi.org/10.1016/j.imavis.2013.12.002

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. Boston, MA. Retrieved from http://www.rstudio.com/

Salim, A., Allen, M., Mariki, K., Masoy, K., & Liana, J. (2023). *Understanding how the use of AI decision support tools affect critical thinking and over-reliance on technology by drug dispensers in Tanzania*. Cornell University. https://doi.org/10.48550/arXiv.2302.09487

Strathie, A., & McNeill, A. (2016). Facial wipes don't wash: Facial image comparison by video superimposition reduces the accuracy of face matching decisions. *Applied Cognitive Psychology, 30*(4), 504–513. https://doi.org/10.1002/acp.3218

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics, 6*, 497-516). Boston, MA: Pearson.

Trinh, A., Dunn, J. D., & White, D. (2022). Verifying unfamiliar identities: Effects of processing name and face information in the same identity-matching task. *Cognitive Research: Principles and Implications, 7*(1), 92–92. https://doi.org/10.1186/s41235-022-00441-2

Weatherford, D. R., Roberson, D., & Erickson, W. B. (2021). When experience does not promote expertise: Security professionals fail to detect low prevalence fake IDs. *Cognitive Research: Principles and Implications, 6*(1), 1-27. https://doi.org/10.1186/s41235-021-00288-z

Weger, Loughnan, S., Sharma, D., & Gonidis, L. (2015). Virtually compliant: Immersive video gaming increases conformity to false computer judgments. *Psychonomic Bulletin & Review, 22*(4), 1111–1116. https://doi.org/10.3758/s13423-014-0778-z

White, D., Guilbert, D., Varela, V., Jenkins, R., & Burton, A. M. (2022). GFMT2: A psychometric measure of face matching ability. *Behaviour research methods, 54*(1), 252–260. https://doi.org/10.3758/s13428-021-01638-x

White, D., Kemp, R., Jenkins, R., Matheson, M., & Burton, A. (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), 62-88. https://doi.org/10.1098/rspb.2015.1292

Wirth, B., & Carbon, C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied, 23*(2), 138– 157. https://doi.org/10.1037/xap0000114