Simulated Automated Facial Recognition Systems as Decision-Aids in Forensic Face Matching Tasks

*Daniel J. Carragher[1, 2] & Peter J. B. Hancock[1]

[1]Psychology

Faculty of Natural Sciences

University of Stirling

Scotland, United Kingdom

[2]School of Psychology

Faculty of Health and Medical Sciences

University of Adelaide

Adelaide, Australia

*****THIS MANUSCRIPT HAS BEEN ACCEPTED FOR PUBLICATION IN THE JOURNAL OF EXPERIMENTAL PSYCHOLOGY: GENERAL (09/09/2022). THIS IS THE AUTHOR ACCEPTED VERSION OF THIS PAPER*****

Total Word Count: 12,100 approx.

*Corresponding Author:
Daniel J. Carragher
School of Psychology
Faculty of Health and Medical Sciences
University of Adelaide
Adelaide, South Australia, 5000
daniel.carragher@adelaide.edu.au

Author note: Prior to data collection, we pre-registered the aims, hypotheses, design, and analyses for each experiment on the OSF. The datasets generated and analysed in the main text and supplementary materials are also available in the same OSF repository [https://osf.io/d4vkm/]. These data were presented at a meeting of the Experimental Psychology Society, UK, July 13-15, 2022.

Abstract

Automated Facial Recognition Systems (AFRS) are used by governments, law enforcement agencies and private businesses to verify the identity of individuals. While previous research has compared the performance of AFRS and humans on tasks of one-to-one face matching, little is known about how effectively human operators can use these AFRS as decision-aids. Our aim was to investigate how the prior decision from an AFRS affects human performance on a face matching task, and to establish whether human oversight of AFRS decisions can lead to collaborative performance gains for the human-algorithm team. The identification decisions from our simulated AFRS were informed by the performance of a real, state-of-the-art, Deep Convolutional Neural Network (DCNN) AFRS on the same task. Across five pre-registered experiments, human operators used the decisions from highly accurate AFRS (>90%) to improve their own face matching performance compared to baseline (sensitivity gain: Cohen's d = 0.71-1.28; overall accuracy gain: d = 0.73-1.46). Yet, despite this improvement, AFRS-aided human performance consistently failed to reach the level that the AFRS achieved alone. Even when the AFRS erred only on the face pairs with the highest human accuracy (>89%), participants often failed to correct the system's errors, while also overruling many correct decisions, raising questions about the conditions under which human oversight might enhance AFRS operation. Overall, these data demonstrate that the human operator is a limiting factor in this simple model of human-AFRS teaming. These findings have implications for the "human-in-the-loop" approach to AFRS oversight in forensic face matching scenarios.

Keywords: human-algorithm teaming; face recognition; automation; verification; collaborative decision-making

Introduction

Facial appearance is commonly used for identification, including when buying age-restricted products, opening a bank account, or crossing international borders. Identification often occurs as a one-to-one face matching task, where an observer must decide whether a form of photographic identification (e.g., passport, driver's licence) matches the identity of the person presenting it for inspection. Despite the apparent simplicity of this task, matching unfamiliar faces (i.e., those that have not been seen before) is surprisingly difficult (Bruce et al., 1999; Hancock et al., 2000; Kemp et al., 1997). Error rates of approximately 10-20% are often reported in face matching tasks with high quality images captured under favourable conditions (Burton et al., 2010), while error rates above 30% are common in tasks with image variability that is more representative of applied scenarios (Carragher & Hancock, 2020; Fysh & Bindemann, 2018b). But there are also considerable differences in face matching ability, such that some individuals struggle to achieve chance performance, while others perform with near perfect accuracy (Bobak et al., 2016; Burton et al., 2010). These individual differences persist among professionals who regularly perform face matching (Weatherford et al., 2021; White et al., 2014; Wirth & Carbon, 2017). As a group, passport issuance officers recorded the same average level of unfamiliar face matching performance as an untrained participant sample; however, the accuracy of individual officers ranged from 57-95% correct (White et al., 2014).

The fallibility of human face matching ability has spurred the development of Automated Facial Recognition Systems (AFRS). In just seconds, these systems can compare a single probe face to another face to verify an individual's identity (e.g., comparing a traveller to their passport image), or to an entire database of known faces (e.g., comparing an image from CCTV to a database of known offenders). While the computational architecture underlying these AFRS is remarkably complex (Noyes & Hill, 2021), simply stated, these

systems must find a face in the submitted image (either an image file or live-feed video) and then process it to output a simple vector of typically 128 - 512 numbers. In verification tasks, the vectors from the two specified images are compared. For recognition, the vector from the single input image is compared to all those stored in the database. In both cases, a threshold value can be set for the AFRS, which determines the level of vector similarity that is required before declaring two images to be an identity "match"; setting a high threshold will reduce false matches (two different people incorrectly judged to be the same) at the cost of excluding some correct matches. Recent technological advancements, including the development of Deep Convolutional Neural Networks (DCNNs), have seen the accuracy of these AFRS increase dramatically over the last two decades (Grother et al., 2021; Phillips & O'Toole, 2014), such that many state-of-the-art systems now outperform all but the very best human observers (Phillips et al., 2018), even on tasks with highly challenging novel stimuli (Carragher & Hancock, 2020; Ngan et al., 2020).

With these advances in accuracy, AFRS are now used to secure highly sensitive infrastructure, including border crossings (Ritchie et al., 2021). Electronic passport gates ("e-Gates") are a form of Automated Border Control (ABC) commonly found in international airports (Fysh & Bindemann, 2018a). These e-Gates contain a document reader, which extracts a stored digital copy of the traveller's passport image, and a camera, which is used to capture images of the traveller's current appearance. AFRS software in the e-Gate then creates and compares vector templates for the two images (passport photograph, current appearance). Templates with a similarity value above a certain threshold are deemed to be an identity match, and the traveller is allowed through the e-Gate (MacLeod & McLindin, 2011).

Despite the high accuracy of many modern AFRS, a human operator is still needed to monitor the performance of the e-Gates for errors or inconclusive judgments (FRONTEX, 2015; Fysh & Bindemann, 2018a), since even highly accurate AFRS can make errors that are

obvious to a human observer (Hancock et al., 2020). Instances in which the e-Gate returns an identity mismatch, or is unable to process an individual, are also referred to a human agent for manual processing (FRONTEX, 2015; MacLeod & McLindin, 2011). Although such "human-in-the-loop" models of AFRS oversight are already in use (FRONTEX, 2015), little is known about the way prior decisions from an AFRS affect the final identification decisions made by the human operator. By investigating how humans use AFRS as a decision-aid in face matching tasks, we hope to shed light on the factors that might affect the efficacy of these human-in-the-loop models of AFRS oversight.

Initial investigations of human-algorithm teaming in face recognition focused on "fusion", a process wherein the independent judgments of humans and AFRS are combined in a weighted average to produce a final identification decision (O'Toole et al., 2007). Much like the "wisdom of the crowds" effect (Galton, 1907), whereby averaging judgments from many observers produces more accurate face matching performance than the best performing individuals (Jeckeln et al., 2018; White et al., 2013), fusing the decisions of humans and algorithms can also result in higher accuracy then either achieves alone (Phillips et al., 2018). These promising results from fusion approaches would seem to suggest that human-AFRS teaming could lead to collaborative performance gains in face matching tasks. However, fusion treats the decisions made by the AFRS and the human as separate events. As such, this approach does not reflect the nature of a sequential decision-making process in which the human is aware of the decision made by the AFRS prior to making their own judgment. This sequential process is potentially consistent with a scenario in which the human operator reviews the decisions made by the AFRS (Fysh & Bindemann, 2018a).

Even though humans (Bruce et al., 1999; Megreya & Burton, 2006) and AFRS (Grother et al., 2019; Grother et al., 2021) can both make face matching errors, it is still possible for a human-algorithm team to achieve a level of performance exceeding that which

either achieves alone (Wickens & Dixon, 2007). In an oversight scenario, this collaborative gain would be realised if the human operator followed the system's correct decisions and overruled its incorrect decisions. Theoretically, optimal collaborative performance could be achieved if the AFRS shared direct evidence information about a decision with the human operator, who after making their own independent decision, was capable of weighing the two judgments by the reliability of each source's past decisions (Bahrami et al., 2010; Robinson & Sorkin, 1985; Sorkin et al., 2001). However, such optimal performance is rarely observed in human-algorithm teams (Bartlett & McCarley, 2017; Boskemper et al., 2021). Instead, human use of automated aides is often typified by *misuse* or *disuse*, wherein the operator over- or under-relies on the decisions from the system (Parasuraman & Riley, 1997), reducing the possible collaborative performance of the pairing (Bahrami et al., 2010; Bartlett & McCarley, 2017).

There are several reasons to think that sub-optimal performance is likely to characterise human use of AFRS in verification tasks. First, although an individual's decision confidence relates to their face matching accuracy on a trial-by-trial basis (Stephens et al., 2017), observers only have a moderate insight into their own general face identification abilities (Bobak et al., 2019; Matsuyoshi & Watanabe, 2021; Zhou & Jenkins, 2020), which could contribute to inappropriate use of the AFRS, whether by misuse or disuse (Lee & Moray, 1994; Parasuraman & Riley, 1997). Second, human performance has already been shown to be a limiting factor in AFRS assisted one-to-many face matching tasks (Heyer et al., 2018; White, Dunn, et al., 2015). White, Dunn et al. (2015) used a commercial AFRS to return a candidate list of the 8 faces that were most similar to each probe identity from a database containing millions of images. Participants only correctly identified the target from the candidate list in 45% of trials where it was present, while also only correctly rejecting target absent arrays on 45% of trials. Concerningly, both the student sample and the

professional passport review officers - who used AFRS in their daily work - made errors on more than 50% of trials (White, Dunn, et al., 2015). In a similar study, Heyer et al. (2018) reported that both novices and professional facial reviewers made more errors as the number of faces in the candidate list returned by the AFRS increased.

Although these studies offer the first indication that human ability might constrain the potential benefits of human-AFRS teaming (Heyer et al., 2018; White, Dunn, et al., 2015), the task demands of a one-to-many array search differ substantially from those of the one-to-one matching task that is performed in many identity verification scenarios. To the best of our knowledge, only two studies have investigated how human performance on a one-to-one face matching task is influenced by the prior decision of an AFRS. Fysh and Bindemann (2018a) had participants complete a matching task in which each face pair was accompanied by a decision label ("same" or "different") from a fictitious AFRS that was correct on 60% of trials, incorrect on 20% of trials, and "unresolved" on 20% of trials. Human performance was highest on trials that were correctly labelled and fell when the labels were incorrect or unresolved. Likewise, Howard et al. (2020) showed participants 12 face pairs, each with an identity decision from a fictitious AFRS that was correct on 50% of trials. Human ratings of similarity for each face pair shifted toward the decision label, regardless of its accuracy. Both studies concluded that humans were biased toward accepting decisions made by the AFRS, even if erroneous (Fysh & Bindemann, 2018a; Howard et al., 2020).

Yet, both studies suffer limitations that affect their generalisability (Fysh & Bindemann, 2018a; Howard et al., 2020). First, neither measured the unaided face matching performance of the participants. As such, whether, and to what extent, human performance changes when working with an AFRS remains unclear. Second, neither study used a real AFRS; instead, the researchers randomly selected and counterbalanced the pairs of faces that were shown with erroneous decision labels. Although this methodological choice is

appropriate from an experimental perspective, real AFRS will only err on the mismatch pairs they calculate to have the highest vector similarity (or identity matches with the lowest similarity ratings). Third, these fictitious AFRS had uncharacteristically low accuracy, which can reduce the operator's reliance on a decision-aid (Wickens & Dixon, 2007), and limits the generalisation of these findings to operational settings with more accurate systems. For example, the European Border and Coast Guard Agency requires that false acceptance of a mismatch by ABCs occurs on fewer than 0.1% of trials, with false rejections of a true match occurring on less than 5% of trials (FRONTEX, 2015). Finally, neither study informed participants about the exact accuracy of the AFRS, which constrains the ability of the observer to appropriately weigh the AFRS's decisions against their own to achieve optimal performance (Bahrami et al., 2010; Robinson & Sorkin, 1985). Together, these limitations mean that many questions remain about how the prior decision from an AFRS affects human face matching performance, which could have implications for the successful implementation of various models of human-in-the-loop AFRS oversight.

The overarching aim of this project was to investigate how the prior decision from an AFRS affects human performance on a face matching task, and to establish whether human oversight of AFRS decisions in a simplified paradigm can lead to performance gains for the collaborative pair. Importantly, and in contrast to previous studies, the decisions given by the simulated AFRS in each experiment were informed by the performance of a real, state-of-the-art, DCNN AFRS on the same face matching task (see Carragher & Hancock, 2020). Although two previous studies have shown that human decisions are biased toward those made by a fictitious AFRS (Fysh & Bindemann, 2018a; Howard et al., 2020), many basic questions remain about the collaborative performance of these human-algorithm teams. Across five pre-registered experiments, we investigated whether it is possible for human operators to improve their own face matching performance when using a simulated AFRS as

a decision-aid, and tested whether this level of aided performance exceeds that of the AFRS alone. We also examined whether the efficacy of this arrangement varies depending on how the AFRS communicates decisions (Experiment 1b), whether participants know accuracy of the AFRS (Experiment 2b), or the difficulty of the trials (high vs. low human accuracy) that the AFRS errs on (Experiment 3). A summary of the conditions in each experiment is reported in Table 1. These experiments were designed to further our limited understanding of collaborative human-algorithm teaming in complex identity verification scenarios.

Table 1

The characteristics of each AFRS across all five experiments ("Expt."). "Decision Type" shows whether the AFRS gave a binary identification decision (same, different) or a decision supplemented with a similarity rating (0.00-1.00). "System Accuracy" indicates whether the participants were told the exact accuracy of the AFRS prior to the task (known, unknown). "Error Trials" describes how the trials that were errors for the AFRS were selected (DCNN similarity ratings, human accuracy). Overall Accuracy and d′ show the level of performance each AFRS achieved alone.

		AFRS Characteristics				
Expt.	**Aid Condition**	**Decision Type**	**System Accuracy**	**Error Trials**	**Overall Accuracy**	**d′**
1a	AFRS	Binary	Known	DCNN Similarity	97.6%	3.962
	Control	-	-	-	-	-
1b	AFRS	Similarity	Known	DCNN Similarity	97.6%	3.962
	Control	-	-	-	-	-
2a	AFRS93	Binary	Known	DCNN Similarity	92.9%	2.930
	AFRS55	Binary	Known	DCNN Similarity	54.8%	0.239
	Control	-	-	-	-	-
2b	AFRS93	Binary	Unknown	DCNN Similarity	92.9%	2.930
	AFRS55	Binary	Unknown	DCNN Similarity	54.8%	0.239
	Control	-	-	-	-	-
3	AFRS-High	Binary	Known	High Human Accuracy	90.5%	2.618
	AFRS-Low	Binary	Known	Low Human Accuracy	90.5%	2.618
	Control	-	-	-	-	-

General Method

All five experiments (1a, 1b, 2a, 2b, 3) rely on the same basic methodology. We outline this general method here and highlight any differences in the experiments as they are reported in turn.

Transparency and Openness

We report all manipulations and measures, participant exclusions, as well as how we determined the sample size in each experiment. The aims, hypotheses, design, and analyses for each experiment were pre-registered prior to data collection. Exploratory analyses are clearly identified in each results section. All analysed data are available online. These pre-registrations and data files can be found on the Open Science Framework (OSF) [https://osf.io/d4vkm/]. All statistical analyses were performed in JASP 0.14.0 (JASP Team, 2020).

Participants

All participants were recruited online through *Prolific (https://prolific.co/)*. Participants were at least 18 years old, reported living in the UK, and met our minimum Prolific experience criteria (had completed at least 5 previous experiments with an accepted completion rate above 90%, but had not participated in any other study from our laboratory).

To maintain data integrity, we applied pre-registered exclusion criteria to these data prior to analysis. All data were excluded from participants who: completed the task too quickly (< 8 minutes) or slowly (> 60 minutes); did not understand task instructions (indicated by baseline d′ < 0); failed to give the correct response to both attention check trials (see below); had missing data (or failed to complete the task); or started the face matching task multiple times[1]. Finally, in Experiment 2b we pre-registered an additional criterion (that we also applied retroactively to Experiments 1a and 2a), to exclude all data from participants who gave a response to one of two end-of-task questions that was either inconsistent with their actual experimental condition (e.g., control condition participants incorrectly reporting that they were in the AFRS condition), or indicated that they did not recall the stated accuracy of the AFRS (from a multiple choice question with two options). We only report the

[1] See the pre-registration of Experiment 1a for further detail.

demographics of the final samples in the main text. Participant exclusions for each experiment are reported in the supplementary materials.

Ethics

The General University Ethics Panel at the University of Stirling approved this research. All participants gave informed consent before starting an experiment, were debriefed on completion, and received £2.50 for their time.

Design

Each experiment had a mixed measures design. Participants were randomly allocated to an Aid Condition for the duration of the experiment (between-participants). All participants completed a face matching task (see below), which was presented in two phases (within-participants). In the baseline phase, all participants completed the face matching task alone. In the test phase, participants in the AFRS condition(s) were shown the identification decision from a simulated AFRS prior to making their own response, whereas those in the control condition did not receive any assistance. Thus, each experiment consisted of the within-participants factor of *Task Phase* (baseline, test) and the between-participants factor of *Aid Condition* (AFRS, control).

Expertise in Facial Comparison Test

Participants completed the Expertise in Facial Comparison Test (EFCT; White et al., 2015), which consists of images from *The Good, The Bad, and The Ugly* stimulus set (Phillips, Beveridge, et al., 2011). This image set contains multiple images of the same individuals, which were collected in unconstrained naturalistic settings on different days, ensuring transient characteristics (e.g., clothing, hairstyle) do not cue identity (see Figure 1). The EFCT consists of face pairs with high error rates among computer algorithms (of the time) and human observers (O'Toole et al., 2012; White, Phillips, et al., 2015). The test has 168 trials and includes both male and female face pairs.

We divided the EFCT into two sets (A, B) of equal difficulty (Carragher et al., 2022; White, Phillips, et al., 2015), which each had 42 match pairs and 42 mismatch pairs. The presentation of each set (A, B) in the baseline or test phase of the experiment was counterbalanced between participants. Trial order was randomised within each set. Custom written code rotated each face to align the eyes horizontally in the centre of the image. The stimuli were presented in colour. Each image was 252 x 357 px in size (approximately 8.0 x 11.5 cm when presented though Qualtrics on a 23″ 1920 x 1080 px monitor).

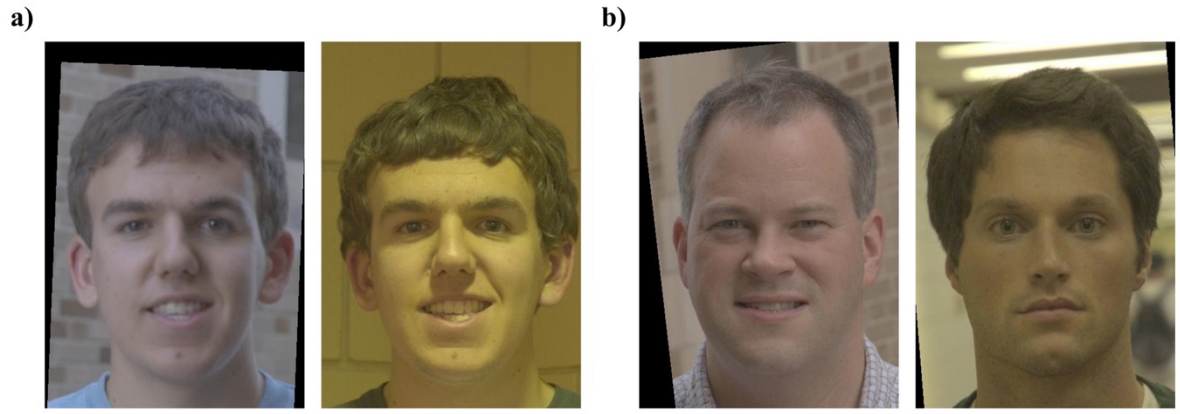

Figure 1. Examples of **a)** match and **b)** mismatch trials from the EFCT. To preserve the integrity of the EFCT, the two pairs shown here are among those with the highest accuracy for human observers.

Automated Facial Recognition Systems

Like the two previous studies in this field (Fysh & Bindemann, 2018a; Howard et al., 2020), we chose to simulate the AFRS in each experiment, which allowed greater experimental control over the performance characteristics of the decision-aid. However, our approach advances on the fictitious AFRS in these previous studies because the decisions from our simulated AFRS were based on those of a real, state-of-the-art, DCNN on the EFCT.

Real AFRS. This DCNN was developed by the University of Surrey, which we had access to through the FACER2VM project. This DCNN produces a similarity score that is the cosine of the angle between the 512 element vectors of the two faces being compared. These

similarity scores can range from -1.0 to +1.0, where scores of +0.40 and above are indicative of a "same" identification decision. We have previously shown that the classification accuracy of the FACER2VM DCNN exceeds that of three commercial DCNNs on the Stirling Famous Face Matching Task (Carragher & Hancock, 2020). Further information about this DCNN can be found in our previous work (Carragher & Hancock, 2020).

Simulated AFRS. The real FACER2VM DCNN correctly classified all 168 pairs in the EFCT (see supplementary materials). However, to address our research questions, each simulated AFRS needed to give several incorrect decisions throughout the task. As such, we selected the match/mismatch pairs in each EFCT set (A, B) that were closest to the DCNN's decision threshold as those that would be shown as "errors" in each experiment[2]. The exact number of trials with incorrect labels varies in each experiment, depending on the level of AFRS performance required. During the aided test phase for the AFRS condition, the AFRS gave a binary identification decision ("same", "different").

Attention Check

We added two attention check trials to the EFCT so that we could screen the data for inattentive or inauthentic participants (Carragher et al., 2022). These were pairs of famous faces that were obvious identity mismatches (Pair 1: former presidents Barack Obama & Donald Trump; Pair 2: Queen Elizabeth II & Prime Minister Boris Johnson). For the attention check trial in the test phase of each experiment, the AFRS did not offer an identification decision – instead, the system showed that it was "*offline*" to all conditions. Participants who failed to respond "*different*" to both attention check trials were excluded from all analyses.

[2] When the DCNN gave the same similarity rating to multiple face pairs, we selected the pair with the lowest average human accuracy as the error trial. These accuracy values came from the "pre-training" responses to the EFCT reported in the supplementary materials of Carragher et al. (2022).

Procedure

Each experiment was run online using Qualtrics survey software. These experiments could not be accessed on mobile devices. The initial instructions informed participants that they were to complete a face matching task. They were then told of our interest in how people complete these tasks when assisted by state-of-the-art facial recognition computers. All participants were told that in the second half of the experiment, some participants would be able to see the decisions made by a simulated AFRS we had created for the experiment. The participants were told exactly how accurate the decisions from the AFRS would be. Participants were instructed that they would still need to give their own identification decision, and that they should correct the AFRS when they thought it had provided the incorrect answer.

On each trial, two faces were presented simultaneously. Participants responded to the question "*Do these photographs show the same person, or two different people?*" by clicking either "*same*" or "*different*". The faces remained on screen until a response was made. After completing the first half of the EFCT on their own (baseline), participants could take a short break. Participants were then reminded that some respondents would see the identity decision made by our simulated AFRS for each pair in the second half of the task and were told again exactly how accurate these decisions would be.

The second half of the task (test) only differed to the first (baseline) in that below the main prompt question there was a new line that read "*Computer says: …*". For participants in the AFRS condition(s), the AFRS gave a binary identification judgment ("*same*" or "*different*") to each pair. For participants in the control condition, the AFRS always gave the response "*offline*" (i.e., they received no assistance). All participants still gave their own identity judgment for each pair.

Following the face matching task, participants were asked several exploratory questions about their beliefs regarding the accuracy of humans and AFRS on face matching tasks. These exploratory data are not reported in the current paper; instead, they were included to inform our future research. Across all five experiments, participants took an average of 18 minutes (*SD* = 8.1) to finish the task.

Analysis

Responses to each trial were recorded as hits (correctly responding "same" on a match trial) and false alarms (incorrectly responding "same" on a mismatch trial), which were used to calculate the signal detection measures (Green & Swets, 1966) of *sensitivity* (d′, "dee-prime") and *criterion* using the formulae in Stanislaw and Todorov (1999). Sensitivity describes how well participants can distinguish identity matches from mismatches, independent of response bias. With a minor adjustment for extreme performance (Stanislaw & Todorov, 1999), the maximum possible value of d′ in each half of the EFCT is 4.52. A d′ of 0 indicates chance performance. As per our pre-registrations, we consider sensitivity to be our primary measure of performance. However, we also report a complete secondary analysis of overall accuracy to provide a description of performance that is more familiar to all readers. In the supplementary materials, we analyse accuracy for match and mismatch trials separately (Megreya & Burton, 2007).

Criterion is a measure of response bias, which shows whether participants tended to respond "*same*" or "*different*" across trial types (Macmillan & Creelman, 2004). As such, criterion is not a measure of performance or ability; rather, criterion offers a descriptive measure of response strategy. Negative values indicate a liberal response bias (responding "same"), whereas positive values signal a conservative response bias (responding "different"). To foreshadow the results from all five experiments, participants in all conditions consistently showed a conservative response bias at baseline, which weakened –

but often remained significant – in the later test phase. This liberal response bias drift is consistent with normal response behaviours in long face matching tasks (Alenezi et al., 2015), and occurred when we previously used the EFCT (Carragher et al., 2022), indicating that it is completely unrelated to our AFRS manipulation. For brevity, the pre-registered one-sample *t*-tests comparing criterion to neutral responding for each condition are reported in the supplementary materials for each experiment.

As per our pre-registrations, we have supplemented frequentist *t*-tests (one- and paired-samples) with equivalent Bayesian *t*-tests. For consistency, all reported Bayes factors test for evidence in favour of the alternative hypothesis (BF_{10}). The following classification scheme (JASP Team, 2020) can be used to characterise the strength of our Bayes factors (Goss-Sampson et al., 2020); Bayes factors of 1-3 provide anecdotal evidence in favour of the *alternative* hypothesis, while factors of 3-10, 10-30, 30-100 and >100 provide moderate, strong, very strong, and decisive evidence, respectively. Conversely, values between 1.00-0.33 provide anecdotal evidence in favour of the *null* hypothesis, while factors of 0.33-0.10, 0.10-0.033, 0.033-0.010, and <0.010 provide moderate, strong, very strong, and decisive evidence, respectively. All Bayesian analyses use default priors (JASP Team, 2020).

Regardless of measure (d′, overall accuracy, criterion), each mixed measures analysis of variance (ANOVA) reported in Experiment 1 (a, b) has Task Phase (baseline, test) as a within-participants factor and Aid Condition (AFRS, control) as a between-participants factor. In Experiment 2 (a, b) and Experiment 3, Aid Condition is a between-participants factor with 3 levels. The interaction between Task Phase and Aid Condition is the key comparison in each ANOVA. Significant interactions are interpreted using simple main effects analysis.

Experiment 1a

We began by investigating how the prior decisions from a highly accurate AFRS would influence human face matching performance. Participants were randomised into two between-subjects Aid Conditions (AFRS, control). The simulated AFRS gave correct responses to 41/42 match trials and 41/42 mismatch trials, for an overall accuracy of 97.6% (d′ = 3.962).

We expected a significant interaction between *Aid Condition* (AFRS, control) and *Task Phase* (baseline, test), such that the performance of the AFRS condition when aided at test would improve compared to baseline. No change was expected for the control condition. However, we also expected that those in the AFRS condition would fail to use the decision-aid in an optimal way (Bartlett & McCarley, 2017), such that their aided sensitivity in the test phase would be significantly lower than that of the AFRS alone.

We also pre-registered testing whether an individual's face matching ability is related to their capacity to use the AFRS decision-aid effectively. We expected to find a correlation between sensitivity at baseline and the change to sensitivity in the aided task; however, because this was an exploratory analysis, we offered no formal prediction regarding the direction of this relationship. This analysis could reveal whether certain individuals are suited to roles that use AFRS as a decision-aid, which might have consequences for personnel selection in applied settings.

In addition to examining human performance, we also investigated how the accuracy for each face pair changed in the aided test phase for the AFRS condition. Across counterbalances the AFRS gave incorrect responses to 4 of 168 trials (i.e., one match and one mismatch error in both Set A and Set B). We expected that accuracy for these four error trials would be lower for the AFRS condition in the aided test phase than at baseline, whereas accuracy for the correctly labelled trials should increase. Because there were so few errors,

we simply report the descriptive statistics for the change in accuracy (accuracy at test minus accuracy at baseline) for pairs that were shown with correct and incorrect AFRS labels. To infer meaningful differences, the changes in accuracy that occur in the AFRS aid condition should exceed those that occur in the control condition.

Method

Sample Size

Experiment 1 (a, b) has a 2 x 2 mixed measures design. However, we could not identify an appropriate prior effect size for an interaction in this paradigm to use in a power analysis. Instead, we conducted our power analysis using the arbitrarily selected medium-to-large effect size of $\eta_p^2 = 0.09$. Our *a priori* power analysis (G*Power; Faul, Erdfelder, Lang, & Buchner, 2007) indicated that 82 participants (total) were required to achieve 80% power to detect an interaction effect of $\eta_p^2 = 0.09$ in a mixed-measures ANOVA with a 2 level within-participants factor (*Task Phase*) and a 2 level between-participants factor (*Aid Condition*) at an alpha of $\alpha = .05$. To account for participant exclusions, we aimed to recruit 55 participants to each Aid Condition, so that data from approximately 50 participants would remain in each condition for the final analysis. This was also true of Experiment 1b.

Participants

The final sample consisted of 53 participants in the AFRS condition ($M = 33.0$, $SD = 11.7$, 30 females, 23 males) and 46 participants in the control condition ($M = 29.6$, $SD = 9.9$, 32 females, 13 males, 1 other).

Results

Participant Performance

Sensitivity

The main effects of Task Phase, $F(1, 97) = 31.35$, $p < .001$, $\eta_p^2 = .24$, and Aid Condition, $F(1, 97) = 8.71$, $p = .004$, $\eta_p^2 = .08$, were significant, as was their interaction, $F(1,$

97) = 23.93, $p < .001$, $\eta_p^2 = .20$ (see Figure 2a). Simple main effects analysis revealed that sensitivity increased in the test phase for the AFRS condition, while there was no change for the control condition (see Table 2). Despite this improvement, a one-sample *t*-test revealed that the aided sensitivity of the AFRS condition at test was significantly lower than that of the AFRS alone (d′ = 3.962), $t(52) = -12.92$, 95%CI of mean difference [-1.88, -1.37], $p < .001$, $d = -1.77$, $BF_{10} = 7.55e+14$.

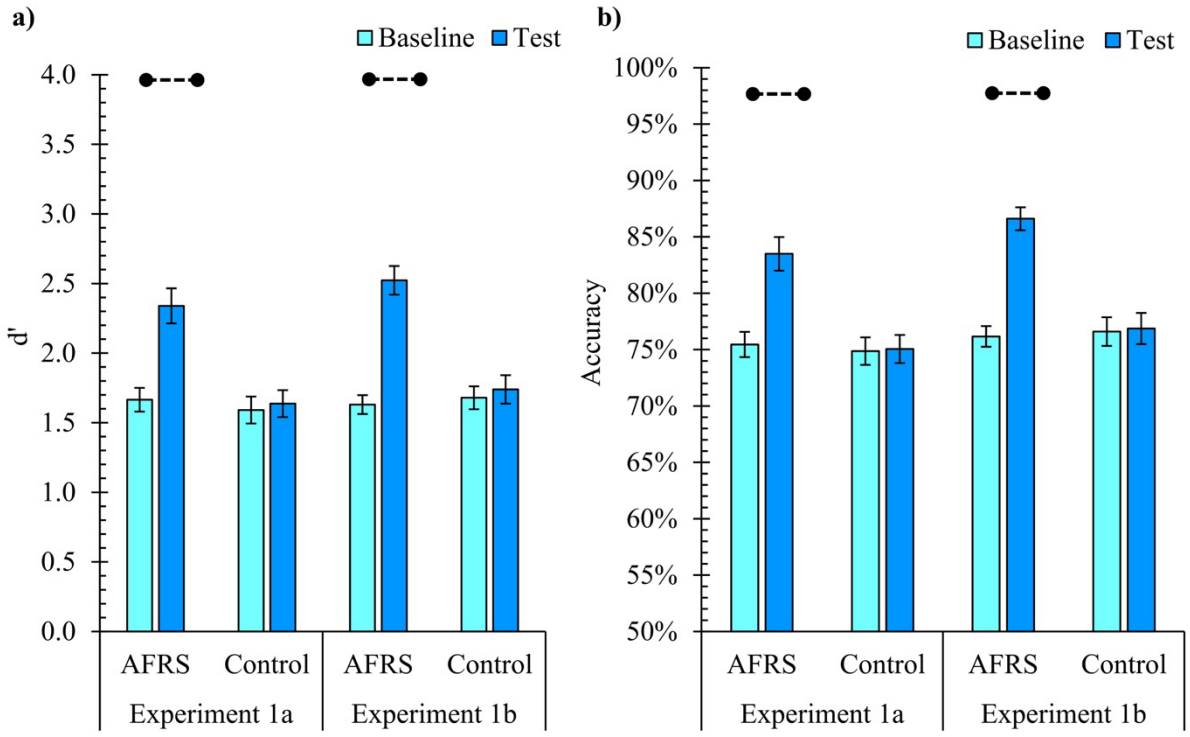

Figure 2. a) Sensitivity and **b)** Overall Accuracy of each Aid Condition (AFRS, control) in the Task Phases (baseline, test) of Experiment 1a (binary decisions) and 1b (similarity decisions). The dashed bars show the performance of the AFRS alone. All error bars show the standard error of the mean (SEM).

Is Face Matching Ability related to Effective AFRS use?

We investigated whether an individual's face matching ability was related to their capacity to use the AFRS decision-aid effectively, as measured by the change to their sensitivity in the aided test phase (test phase d′ minus baseline d′). This relationship was non-significant, $r(52) = -0.11$, 95%CI[-0.37, 0.17], $p = .448$. This result is supported by further analyses reported in the supplementary materials for Experiment 1b.

Table 2

Simple main effects analyses for Experiment 1a (binary decisions) and 1b (similarity decisions), showing sensitivity and overall accuracy for each Aid Condition (AFRS, control). 95%CI are given for the mean difference between the Baseline and Test task phases.

Measure	Aid	Baseline	Test	*t*	*df*	95%CI	*p*	*d*	BF_{10}
Experiment 1a									
Sensitivity	AFRS	1.67 (0.62)	2.34 (0.91)	6.61	52	0.47, 0.88	< .001*	0.91	635,293
	Control	1.59 (0.66)	1.64 (0.66)	0.63	45	-0.10, 0.19	.533	0.09	0.193
Accuracy	AFRS	75.45 (8.14)	83.49 (10.82)	6.58	52	5.59, 10.50	< .001*	0.90	556,137
	Control	74.87 (8.27)	75.05 (8.48)	0.18	45	-1.86, 2.23	.859	0.03	0.162
Experiment 1b									
Sensitivity	AFRS	1.63 (0.47)	2.52 (0.71)	8.85	47	0.69, 1.10	< .001*	1.28	6.53e+8
	Control	1.68 (0.56)	1.74 (0.70)	0.80	46	-0.09, 0.21	.426	0.12	0.215
Accuracy	AFRS	76.17 (6.35)	86.61 (7.03)	10.10	47	8.36, 12.52	< .001*	1.46	3.36e+10
	Control	76.60 (8.72)	76.87 (9.49)	0.29	46	-1.69, 2.25	.777	0.04	0.165

Overall Accuracy

The main effect of Task Phase was significant, $F(1, 97) = 25.82$, $p < .001$, $\eta_p^2 = .21$, as was the main effect of Aid Condition, $F(1, 97) = 7.63$, $p = .007$, $\eta_p^2 = .07$, and their interaction, $F(1, 97) = 23.59$, $p < .001$, $\eta_p^2 = .20$ (see Figure 2b). Overall accuracy increased in the test phase for the AFRS condition. There was no change for the control condition (see Table 2).

Criterion

The main effect of Task Phase was significant, $F(1, 97) = 16.46$, $p < .001$, $\eta_p^2 = .15$. Participants had a larger conservative response bias at baseline ($M = 0.36$, $SD = 0.44$) than at test ($M = 0.20$, $SD = 0.50$). This conservative response bias indicates that participants tended to respond "different" across all pairs. The main effect of Aid Condition was non-significant, $F(1, 97) = 0.71$, $p = .400$, $\eta_p^2 = .01$, as was the interaction, $F(1, 97) = 0.01$, $p = .932$, $\eta_p^2 = .00$.

Image Pair Analysis

AFRS Accuracy Change

To complete our analysis, we investigated the change in accuracy for each image pair when shown with a decision label (correct or incorrect) from the AFRS. We have calculated "change in accuracy" as a simple difference score (i.e., the change from 55% to 65% is a 10%

change). Among the AFRS condition, the average change in accuracy for the correctly labelled pairs was positive, while the average change for error pairs was negative, suggesting that participants generally followed the decision from the AFRS (see Table 3). Curiously, however, only once did the largest negative change across all 84 trials (82 correctly labelled, 2 incorrectly labelled) in an identity condition (match, mismatch) come from an error pair (mismatch error, Set A), suggesting that factors other than the AFRS decision label might influence accuracy. We investigate this possibility in Experiment 3. Nonetheless, the average change values for both the correctly and incorrectly labelled trials in the AFRS condition were more extreme than the corresponding values in the control condition, indicating participants tended to follow the decisions of the AFRS, correct or otherwise. This conclusion is supported by the rank change position for the error pairs, which were more extreme when shown with the incorrect label in the AFRS condition than in the control condition.

Table 3

Difference in accuracy from baseline (%) in the test phase, reported separately for the 82 correctly labelled trials and 2 error pairs (one per image set: A, B) in each identity condition (match, mismatch). "*Max*" and "*Min*" show the maximum and minimum change values from the 82 trials shown with the correct decision label from the AFRS. "*Set A (Rank All)*" and "*Set B (Rank All)*" show the change in accuracy for the single error pair in each EFCT set (A, B), along with the rank order of that change out of all 84 match/mismatch trials in the EFCT (with 1 being the largest increase in accuracy, and 84 the biggest decrease).

Aid Cond.	Pair	Correct (n = 82)			Error (n = 2)		
		Average Change (SD)	Max	Min	Average Change (SD)	Set A Error (Rank All)	Set B Error (Rank All)
AFRS	Match	14.20 (12.19)	54.57	-12.46	-4.97 (9.43)	-11.64 (83)	1.69 (69)
	Mismatch	3.12 (8.94)	31.75	-16.92	-14.83 (5.96)	-19.05 (84)	-10.62 (78)
Control	Match	2.53 (11.72)	33.36	-26.61	5.23 (6.66)	9.94 (23)	0.52 (46)
	Mismatch	-2.86 (10.54)	33.65	-26.78	-2.60 (0.72)	-3.11 (45)	-2.09 (41)

Exploratory Image Analysis

The average accuracy for each face pair at baseline was related to the change in accuracy that pair experienced when shown at test with the correct label from the AFRS,

$r(163) = -0.70$, 95%CI[-0.77, -0.61], $p < .001$ (see Figure 3). Although this relationship persisted among the control condition, $r(163) = -0.36$, 95%CI[-0.48, -0.21], $p < .001$, Fisher's r-to-z transformation confirmed that the two correlations differed significantly, $z = 4.40$, $p < .001$ (Weiss, 2011). Thus, the difficulty of the face pair is related to the benefit from the AFRS, over and above any change that can be attributed to factors such as practice effects or mean reversion. Perhaps unsurprisingly, participants were most likely to benefit from the AFRS when it gave the correct response to the most difficult pairs.

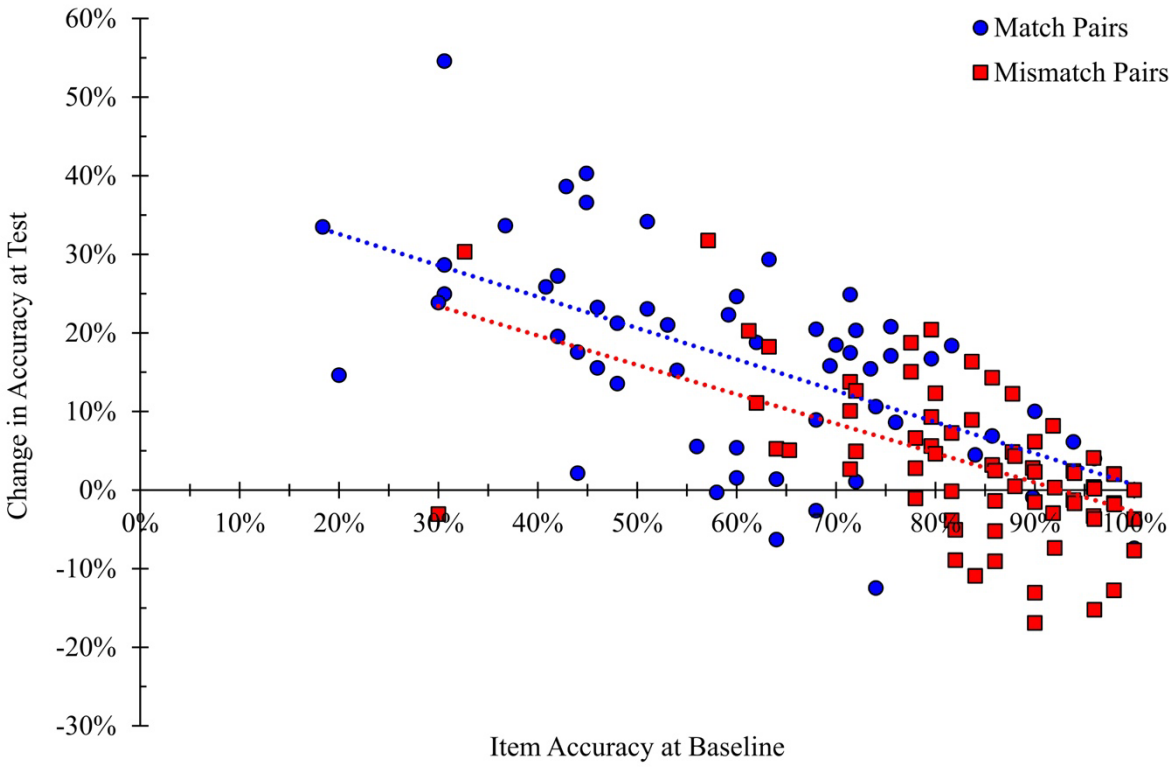

Figure 3. The correlation between item accuracy at baseline and the change in accuracy when the AFRS gave the correct answer in the test phase. For illustrative purposes, this correlation is plotted separately for match (blue circles) and mismatch (red squares) trials.

Discussion

Participants significantly improved their face matching performance compared to baseline when aided by the highly accurate AFRS at test. However, this level of aided performance was still far below that of the AFRS alone, indicating that collaborative performance was sub-optimal (Bartlett & McCarley, 2017). Failing to reach the level of

performance achieved by the AFRS alone demonstrates that participants often overruled correct decisions from the AFRS, even though they were truthfully told the system would be accurate on 97.6% of trials. Thus, even though participants improved their own performance by following the decisions of the AFRS (Fysh & Bindemann, 2018a; Howard et al., 2020), our data suggest that this bias is somewhat incomplete, as further improvement is possible with greater reliance on the AFRS.

Yet, there would be little point to human-algorithm teaming if collaborative performance only equaled that which the AFRS could achieve alone. Collaborative performance gains require the human operator to follow the system's correct decisions and overturn incorrect decisions. Yet, our data also show that accuracy fell on trials that were labelled incorrectly by the AFRS, demonstrating that participants often endorsed these erroneous decisions. When considered alongside the tendency to overturn correct decisions, these findings suggest that participants have limited insight into the accuracy of the AFRS on any one trial. Finally, the baseline ability (d′) of the participants did not predict their ability to use the AFRS effectively, but an analysis of item accuracy showed that correct AFRS decisions were of the greatest benefit for the most difficult face pairs. While this result is consistent with the notion that participants might defer to the AFRS on trials they find particularly challenging, it might also simply reflect the fact that more improvement was possible for the most difficult face pairs.

Experiment 1b

In Experiment 1a, the AFRS communicated identity judgments as binary decisions. While AFRS might ultimately give a binary identification decision, these judgments are based on the underlying "similarity" score the AFRS calculates between the two images. In Experiment 1b, we examined whether having the AFRS report a similarity value for each pair, in addition to the binary decision, would influence AFRS-aided face matching

performance. This similarity value might be interpreted by the observer as an indication of the AFRS's confidence in each judgment, which is the type of additional information that can be used to weigh decisions appropriately in order to achieve optimal collaborative performance (Bahrami et al., 2010; Sorkin et al., 2001).

We expected to replicate the pattern of results in Experiment 1a, and as such, offer identical hypotheses for the analysis of Experiment 1b. Crucially, we also pre-registered a direct comparison between Experiments 1a and 1b, to test the effect of *Decision-Type* (between-participants; binary, similarity). We expected to find a significant three-way interaction, such that the increase in sensitivity for the AFRS condition at test would be larger in Experiment 1a than 1b. Although communicating direct evidence information should improve aided decision-making, Experiment 1a showed that participants had a tendency to overrule the AFRS. We speculated that a smaller increase in AFRS-aided accuracy might occur in Experiment 1b because participants could be more willing to overrule the AFRS on pairs with similarity values closer to threshold. Because the AFRS still had an overall accuracy of 97.6%, overruling the system more often would lead to lower aided performance.

Method

Participants

The final sample consisted of 48 participants in the AFRS condition (M = 31.2, SD = 10.2, 34 females, 14 males), and 47 participants in the control condition (M = 34.6, SD = 12.7, 37 females, 10 males).

Design

The only change from Experiment 1a was that, in addition to the binary identification decision, the AFRS gave a "similarity value" for each face pair (e.g., "*Computer says: 0.75 (same)*"). These similarity values were based on those from the real DCNN, such that they maintained the true rank order of the face pairs. However, we created a new set of normally

distributed values (M = 0.23, SD = 0.09) for the 41 correctly labelled mismatch trials, before creating a mirrored distribution (1.00 minus the mismatch similarity value) for the match trials, ensuring identical distributions in both identity conditions (match, mismatch). The same similarity values were used in Set A and Set B. The new similarity values ranged from 0.00-1.00, and participants were told that the threshold for the simulated AFRS to declare a match was 0.50. The two "error" trials in each set were given similarity values of 0.45 (match pair declared "different") and 0.55 (mismatch pair declared "same"). Both errors were the closest pair to the decision threshold in each identity condition[3], consistent with the proposition that the "AFRS" was least sure about the classification of these pairs.

Results

Sensitivity

The main effects of Task Phase, $F(1, 93) = 57.37$, $p < .001$, $\eta_p^2 = .38$, and Aid Condition, $F(1, 93) = 11.14$, $p = .001$, $\eta_p^2 = .11$, were significant (see Figure 2a), as was the interaction between the two factors, $F(1, 93) = 43.86$, $p < .001$, $\eta_p^2 = .32$. Sensitivity increased in the aided test phase for the AFRS condition. No change occurred for the control condition (see Table 2). Despite their improvement, the aided sensitivity of the AFRS condition at test was significantly lower than that of the AFRS alone (d′ = 3.962), $t(47) = -14.03$, 95%CI of mean difference [-1.65, -1.23], $p < .001$, $d = -2.03$, $BF_{10} = 2.62e+15$.

Overall Accuracy

The main effect of Task Phase was significant, $F(1, 93) = 56.58$, $p < .001$, $\eta_p^2 = .38$, as was the main effect of Aid Condition, $F(1, 93) = 9.94$, $p = .002$, $\eta_p^2 = .10$, and the interaction between the factors, $F(1, 93) = 50.85$, $p < .001$, $\eta_p^2 = .35$. Overall accuracy

[3] In addition to being equidistant to the 0.50 decision threshold, the similarity values for both error pairs were also 0.03 units away from the closest correctly labelled pair.

increased in the aided test phase for the AFRS condition, whereas no change occurred for the control condition (see Table 2).

Criterion

The main effect of Task Phase was significant, $F(1, 93) = 21.19$, $p < .001$, $\eta_p^2 = .19$, with a larger conservative response bias at baseline ($M = 0.29$, $SD = 0.47$) than at test ($M = 0.13$, $SD = 0.52$). The main effect of Aid Condition was non-significant, $F(1, 93) = 1.49$, $p = .226$, $\eta_p^2 = .02$. The interaction was also non-significant, $F(1, 93) = 0.67$, $p = .414$, $\eta_p^2 = .01$.

Cross Experiment Analysis: The Effect of Decision-Type

A mixed measures ANOVA with *Task Phase* as a within-participants factor, and *Aid Condition* and *Decision-Type* (binary, similarity) as between-participants factors revealed that the effect of Decision-Type was not significant at any level (all F's < 1.68, p's > .196, $\eta_p^2 <$ .01; see supplementary materials for full ANOVA). Therefore, aided performance did not differ depending on whether the AFRS communicated decisions using binary decisions alone or including similarity values.

Discussion

Seeking to improve the aided performance achieved by the human operator, binary identity decisions ("same", "different") from the AFRS were supplemented with a similarity value (0.00-1.00). Although we replicated the improvement in performance seen in Experiment 1a, our cross-experiment analysis revealed no effect of "decision-type" on aided performance. This non-significant result of decision-type is consistent with previous findings in the automation literature about the non-significant effect of cue type on aided decision-making (Bartlett & McCarley, 2019). Overall, Experiment 1 demonstrates that while human operators can improve their face matching performance when assisted by a highly accurate AFRS, this level of aided performance fails to reach that of the AFRS alone.

Experiment 2a

Although we'd hope that AFRS deployed in real world scenarios are highly accurate, like those in Experiment 1 (FRONTEX, 2015), the accuracy of individual systems can vary substantially (Grother et al., 2021). As such, the aim of Experiment 2 was to investigate whether decisions from a low accuracy AFRS would impair human face matching performance. To this end, we added an AFRS with low accuracy (54.8%) to the between-participants factor of Aid Condition. The accuracy of the high performing AFRS was lowered to 92.9% (see below). The two AFRS are designated AFRS93 (high accuracy) and AFRS55 (low accuracy), in reference to their overall accuracy. Participants were randomly allocated to one of three Aid Conditions (AFRS93, AFRS55, control) for the entire experiment.

Since prior work has shown that automated aids with low reliability can impair performance (Wickens & Dixon, 2007), and that participants are biased to follow the decisions from an AFRS (Fysh & Bindemann, 2018a), we expected a significant interaction between Task Phase (baseline, test) and Aid Condition (AFRS93, AFRS55, control), such that sensitivity would increase at test compared to baseline for the high accuracy AFRS93 condition (replicating Experiment 1) but decrease for the AFRS55 condition. No change was expected for the control condition. However, as in Experiment 1, we also predicted that the aided sensitivity of the AFRS93 condition at test would be significantly lower than the AFRS alone ($d' = 2.930$). Conversely, we expected that the aided sensitivity of the AFRS55 condition would exceed that of the AFRS alone ($d' = 0.239$), but only because the average participant had a baseline $d' = 1.63$ in Experiment 1a (see also Bartlett & McCarley, 2021).

Method

Sample Size

Again, we could not identify an appropriate prior effect size for a 3 x 2 mixed measures interaction in this paradigm that we could use in a power analysis for Experiment 2

(a, b) and Experiment 3. As in Experiment 1, we conducted our power analysis using the arbitrarily selected medium-to-large effect size of $\eta_p^2 = 0.09$. This a priori power analysis (Faul et al., 2007) indicated that 102 participants (total) were required to achieve 80% power to detect an interaction effect of $\eta_p^2 = 0.09$ in a mixed-measures ANOVA with a 2 level within-participants factor (Task Phase) and a 3 level between-participants factor (Aid Condition) at an alpha of $\alpha = .05$. Although adequate power could be achieved with 34 participants in each aid condition, we committed to recruiting 55 participants to each condition to be consistent with Experiment 1.

Participants

The final sample consisted of 45 participants in the AFRS93 condition (*M* = 31.5, *SD* = 9.1, 33 females, 11 males, 1 other), 46 in the AFRS55 condition (*M* = 34.2, *SD* = 9.7, 34 females, 12 males) and 45 in the control condition (*M* = 29.0, *SD* = 9.0, 33 females, 11 males, 1 response withheld).

Design

This experiment only differs from Experiment 1a in that there were 2 AFRS conditions. The first, AFRS93, gave correct decisions on 39/42 match and 39/42 mismatch trials (92.9% accuracy, d′ = 2.930), while AFRS55 was accurate on 23/42 match and 23/42 mismatch trials (54.8% accuracy, d′ = 0.239). These levels of performance were determined using the baseline EFCT performance of all participants from Experiment 1a (*n* = 99, Mean d′ = 1.63, SD = 0.63) such that the sensitivity of AFRS93 was 2.05SD above average human performance, whereas AFRS55 was 2.20SD below[4].

[4] An additional exclusion criterion was retroactively applied to data from Experiment 1a (see General Method), after we had already used the original sample to determine the performance of these two AFRS. The original calculation included data from 107 participants (Mean d′ = 1.59, SD = 0.65), such that AFRS93 was 2.06SD above average human sensitivity, whereas AFRS55 was 2.09SD below.

Results

Sensitivity

The main effects of Task Phase, $F(1, 133) = 10.43$, $p = .002$, $\eta_p^2 = .07$, and Aid Condition, $F(2, 133) = 12.64$, $p < .001$, $\eta_p^2 = .16$, were significant (see Figure 4a), as was the interaction between the two factors, $F(2, 133) = 24.61$, $p < .001$, $\eta_p^2 = .27$. Sensitivity increased in the aided test phase for the AFRS93 condition. There was no change at test for the AFRS55 or control conditions (see Table 4).

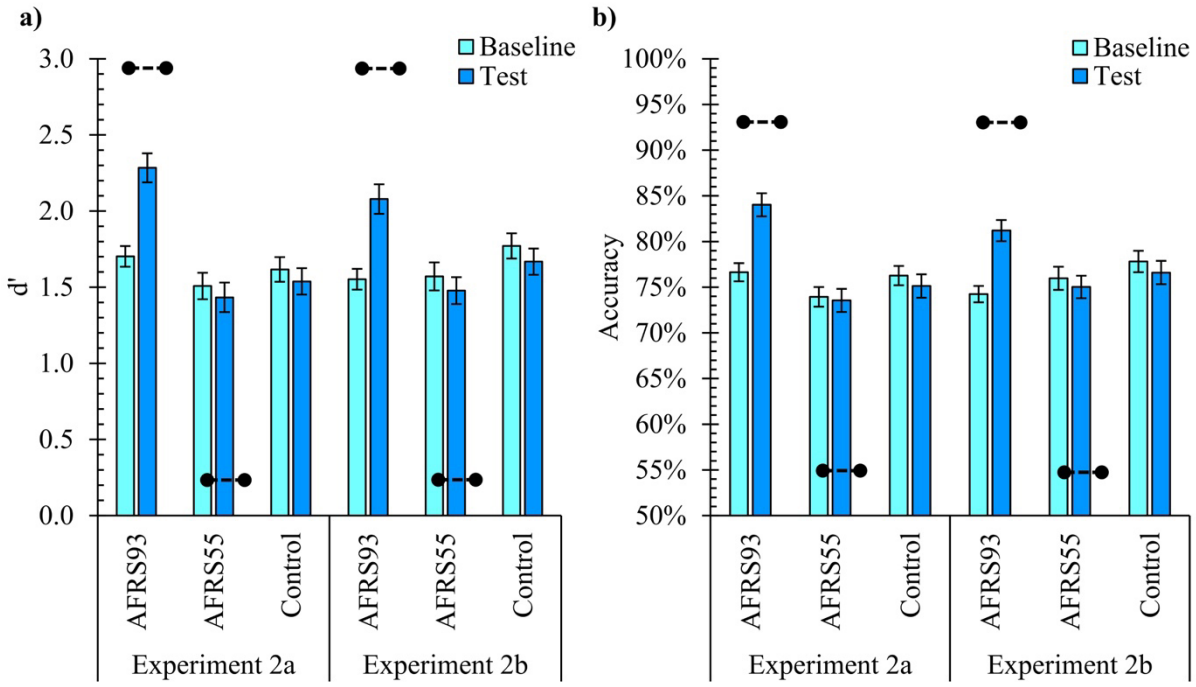

Figure 4. a) Sensitivity and **b)** Overall Accuracy for each Aid Condition (AFRS93, AFRS55, control) in the Task Phases (baseline, test) of Experiment 2a (known accuracy) and 2b (unknown accuracy). The dashed bars show the performance of each AFRS alone. All error bars show the SEM.

As expected, the aided sensitivity of the AFRS93 condition at test was significantly lower than that of the AFRS alone (d′ = 2.930), $t(44) = -6.80$, 95%CI[-0.84, -0.46], $p < .001$, $d = -1.01$, $BF_{10} = 581{,}721$. Conversely, the aided test sensitivity of the AFRS55 condition exceeded their AFRS (d′ = 0.239), $t(45) = 12.29$, 95%CI[1.00, 1.39], $p < .001$, $d = 1.81$, $BF_{10} = 1.13e{+}13$. However, since those in the AFRS55 condition did not improve their sensitivity

from baseline when working with the AFRS, this result is not an example of a collaborative performance gain.

Table 4

Simple main effects analyses for Experiment 2a (known accuracy) and 2b (unknown accuracy), showing sensitivity and overall accuracy for each Aid Condition (AFRS93, AFRS55, control). 95%CI are given for the difference between the Baseline and Test Task Phases.

Measure	Aid	Baseline	Test	*t*	*df*	95%CI	*p*	*d*	BF_{10}
Experiment 2a									
Sensitivity	AFRS93	1.70 (0.46)	2.28 (0.64)	7.30	44	0.42, 0.74	< .001*	1.09	2.94e+6
	AFRS55	1.51 (0.59)	1.43 (0.66)	-0.88	45	-0.25, 0.10	.386	0.13	0.229
	Control	1.62 (0.54)	1.54 (0.58)	-1.28	44	-0.20, 0.05	.208	0.19	0.346
Accuracy	AFRS93	76.64 (6.71)	84.02 (8.51)	6.72	44	5.17, 9.60	< .001*	1.00	453,180
	AFRS55	73.94 (7.25)	73.55 (8.53)	-0.33	45	-2.74, 1.96	.741	0.05	0.169
	Control	76.27 (7.07)	75.13 (8.58)	-1.17	44	-3.11, 0.83	.250	0.17	0.304
Experiment 2b									
Sensitivity	AFRS93	1.55 (0.49)	2.08 (0.70)	5.15	51	0.32, 0.73	< .001*	0.71	4,244
	AFRS55	1.57 (0.65)	1.48 (0.62)	-1.20	48	-0.25, 0.06	.238	0.17	0.304
	Control	1.77 (0.57)	1.67 (0.59)	-1.52	46	-0.24, 0.03	.135	0.22	0.464
Accuracy	AFRS93	74.25 (6.45)	81.20 (8.33)	5.29	51	4.32, 9.60	< .001*	0.73	6,641
	AFRS55	75.97 (8.89)	75.02 (8.67)	-0.88	48	-3.12, 1.22	.384	0.13	0.223
	Control	77.81 (8.04)	76.60 (8.78)	-1.15	46	-3.34, 0.91	.255	0.17	0.295

Overall Accuracy

The main effect of Task Phase was significant, $F(1, 133) = 9.72$, $p = .002$, $\eta_p^2 = .07$, as was the main effect of Aid Condition, $F(2, 133) = 10.87$, $p < .001$, $\eta_p^2 = .14$, and their interaction, $F(2, 133) = 18.83$, $p < .001$, $\eta_p^2 = .22$ (see Figure 4b). Overall accuracy increased at test for the AFRS93 condition. There was no change for the AFRS55 or control conditions (see Table 4).

Criterion

The main effect of Task Phase was significant, $F(1, 133) = 20.25$, $p < .001$, $\eta_p^2 = .13$, with a larger conservative response bias at baseline ($M = 0.28$, $SD = 0.43$) than at test ($M = 0.15$, $SD = 0.43$). The main effect of Aid Condition was non-significant, $F(2, 133) = 2.50$, $p = .086$, $\eta_p^2 = .04$. The interaction was also non-significant, $F(2, 133) = 1.55$, $p = .217$, $\eta_p^2 = .02$.

Discussion

Participants improved their face matching performance compared to baseline when assisted by the highly accurate AFRS93, but failed to outperform the AFRS alone, replicating Experiment 1. Surprisingly, decisions from the low accuracy AFRS55 did not impair participant performance. This is despite previous research suggesting that automated aids with low reliability can impair performance (Wickens & Dixon, 2007). Rather, there was simply no change to the face matching performance of the AFRS55 condition. One possible explanation for this result is that participants decided to give minimal weighting to the decisions from AFRS55 after learning that it would only give the correct response on 54.8% of trials.

Experiment 2b

In each experiment so far, participants were told precisely how accurate their AFRS was before the task. But human operators of real AFRS technology might not be aware of their system's exact accuracy. Without knowing the accuracy of the system, the operator is left to guess how much weight they should give the decisions from the AFRS, which could impair collaborative performance if assessed incorrectly. Here we conduct a direct replication of Experiment 2a, but without telling participants the accuracy of the AFRS before the task. The aim of this experiment was to test how decisions from AFRS of unknown accuracy would influence aided face matching performance.

For the results of Experiment 2b itself, we pre-registered identical hypotheses to those given in Experiment 2a. Crucially, we also pre-registered a cross experiment comparison to investigate whether knowing the accuracy of the AFRS prior to the task would influence aided performance. We expected a significant three-way interaction between Task Phase, Aid Condition, and Accuracy Knowledge (between-participants; known, unknown), such that the increase in sensitivity for the AFRS93 condition would be larger in Experiment 2a than

Experiment 2b, and that there would be a decrease in performance for the AFRS55 condition in Experiment 2b as opposed to no change in Experiment 2a. We expected that these results would arise from participants in the AFRS93 condition under-utilising the decision-aid when not told of its high accuracy before the task, and those in the AFRS55 condition over-relying on the decisions from their aid with undisclosed low accuracy.

Method

Participants

The final sample consisted of 52 participants in the AFRS93 condition (M = 32.5, SD = 11.5, 37 females, 15 males), 49 in the AFRS55 condition (M = 28.0, SD = 7.7, 30 females, 19 males) and 47 in the control condition (M = 32.8, SD = 10.6, 30 females, 17 males).

Design

The only methodological change from Experiment 2a was that participants were not told the exact accuracy of the AFRS before the task. Instead, all participants were simply told that the AFRS would be correct on "most" trials, which was true of both AFRS conditions.

Results

Sensitivity

The main effects of Task Phase, $F(1, 145) = 5.02$, $p = .027$, $\eta_p^2 = .03$, and Aid Condition, $F(2, 145) = 3.98$, $p = .021$, $\eta_p^2 = .05$, were significant (see Figure 4a), as was the interaction between the two factors, $F(2, 145) = 18.40$, $p < .001$, $\eta_p^2 = .20$. Sensitivity increased in the aided test phase for the AFRS93 condition. No change occurred for the AFRS55 or control conditions (see Table 4).

The level of aided performance achieved by the AFRS93 condition at test was significantly lower than that of the AFRS alone (d′ = 2.930), $t(51) = -8.79$, 95%CI[-1.05, -0.66], $p < .001$, $d = -1.22$, $BF_{10} = 1.06e+9$. Once again, the aided sensitivity of the AFRS55

condition exceeded the AFRS alone (d′ = 0.239), $t(48) = 14.05$, 95%CI[1.06, 1.42], $p < .001$, $d = 2.01$, $BF_{10} = 4.20e+15$.

Overall Accuracy

The main effect of Task Phase was significant, $F(1, 145) = 5.62$, $p = .019$, $\eta_p^2 = .04$, but the main effect of Aid Condition was not, $F(2, 145) = 1.33$, $p = .268$, $\eta_p^2 = .02$. The interaction between the two factors was significant, $F(2, 145) = 16.24$, $p < .001$, $\eta_p^2 = .18$ (see Figure 4b). Overall accuracy increased at test for the AFRS93 condition. There was no change for the AFRS55 or control conditions (see Table 4).

Criterion

The main effect of Task Phase was significant, $F(1, 145) = 39.39$, $p < .001$, $\eta_p^2 = .21$, with a larger conservative response bias at baseline ($M = 0.26$, $SD = 0.41$) than at test ($M = 0.08$, $SD = 0.44$). The main effect of Aid Condition was non-significant, $F(2, 145) = 0.56$, $p = .575$, $\eta_p^2 = .01$. The interaction was also non-significant, $F(2, 145) = 1.27$, $p = .283$, $\eta_p^2 = .02$.

Cross Experiment Analysis: The Effect of Accuracy Knowledge

A mixed measures ANOVA with *Task Phase* as a within-participants factor, and *Aid Condition* and *Accuracy Knowledge* (known, unknown) as between-participants factors revealed that the effect of Accuracy Knowledge was not significant at any level (all F's < 2.37, p's > .096, $\eta_p^2 < .02$; see supplementary materials for full ANOVA). Therefore, aided performance did not differ whether the accuracy of the AFRS was known in advance or not.

Discussion

The results of Experiment 2b directly replicate those of Experiment 2a, even though participants were not told the exact accuracy of either AFRS. Thus, the unchanged performance of participants in the AFRS55 condition cannot be attributed to a deliberate strategy to reduce the weighting given to decisions from the AFRS, because the participants

did not know they would be low accuracy. Instead, these results suggest that participants can gauge, at least to some degree, the accuracy of the AFRS they are working with. This conclusion is consistent with previous research showing that human operators come to rely on reliable aids more than unreliable aids, despite not being told of their reliability (Ross et al., 2008). One possibility is that participants gauge the likely accuracy of the AFRS based on the difficulty of the trials the system errs on. An AFRS that makes errors on face pairs that can be correctly resolved by the human operator might come to be seen as "low accuracy" over time because participants are more likely to notice such mistakes.

Experiment 3

Despite differing in accuracy, the AFRS in Experiments 1 and 2 made "errors" on face pairs that were selected according to the similarity values from the real DCNN. However, the correlation between the DCNN's similarity ratings and human accuracy is imperfect. Using human accuracy at baseline for each face pair from Experiments 1a, 2a and 2b (n = 383; Set A = 192, Set B = 191), we find a large positive correlation with the DCNN's similarity ratings for match trials, $r(83) = 0.56$, 95%CI[0.39, 0.69], $p < .001$, and a small negative correlation for mismatch trials, $r(83) = -0.27$, 95%CI[-0.46, -0.06], $p = .013$ (see Figure 5). This result is consistent with Hancock et al. (2020), who also report a larger correlation between human accuracy and DCNN similarity ratings for match trials than mismatch trials.

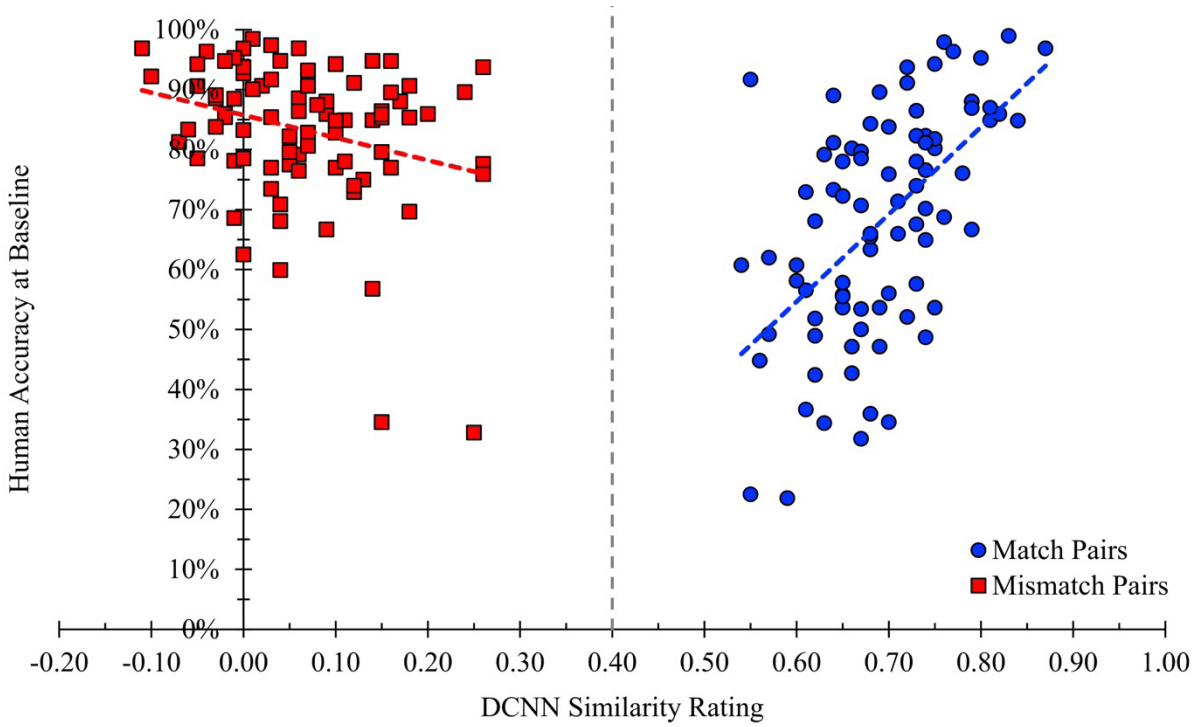

Figure 5. The correlation between human accuracy (%) on the EFCT at baseline (from Experiments 1a, 2a and 2b) and the similarity ratings for each pair from the real DCNN. The DCNN's true decision threshold (0.40) is shown by the vertical grey dashed line.

Despite being correlated across both match and mismatch trials, Figure 5 shows that human accuracy and the DCNN's similarity ratings can differ substantially for some face pairs. For example, there are some pairs that the DCNN gives a similarity rating relatively close to threshold, suggesting that the pair might be challenging due to their similarity, but human accuracy exceeds 90%. However, there are also cases where the DCNN returns a value far from threshold, suggesting that the pair should be easily resolved, but human accuracy is below 50%. Therefore, even though we previously selected error trials based on their proximity to the DCNN's threshold, Figure 5 shows that there is substantial variability in the difficulty of these pairs for human observers. This final experiment tested whether the difficulty of trials (high human accuracy, low human accuracy) that the simulated AFRS errs on can influence the level of aided performance achieved by the human operator.

We created two AFRS conditions. “AFRS-Low” made errors on the face pairs with the lowest human accuracy when shown at baseline (i.e., hard errors), whereas “AFRS-High” made errors on the trials with the highest human accuracy (i.e., easy errors). Importantly, both AFRS (-High, -Low), made errors on 4 match and 4 mismatch trials, giving them the same overall accuracy of 90.5% ($d' = 2.618$).

Because the exploratory image analysis in Experiment 1a revealed that participants benefit most from seeing correct AFRS decisions to the hardest face pairs, we predicted a significant interaction between Task Phase (baseline, test) and Aid Condition (AFRS-High, AFRS-Low, control), such that the increase in sensitivity at test would be larger for the AFRS-High condition than the AFRS-Low condition. This pattern of results would occur if participants overruled the obvious errors made by AFRS-High and followed its correct judgments on all other trials, which included the most difficult face pairs. No change was expected for the control condition.

In contrast to our previous predictions and findings, we expected that the aided test phase sensitivity of the AFRS-High condition would exceed that of the AFRS alone, because participants should correct the AFRS when it errs only on trials that most humans answer correctly. Conversely, we expected that the aided test phase sensitivity achieved by the AFRS-Low condition would be significantly worse than their AFRS, because the participants would be less likely to notice when the AFRS errs on the trials with the lowest human accuracy.

As in Experiment 1a, we also examined the change in accuracy for each face pair when shown with decisions from the AFRS in the aided test phase. Once again, we present the descriptive statistics for the change in accuracy separately for each AFRS decision label condition (correct, incorrect, control). Our overarching prediction was that accuracy would decrease when error pairs were shown with an incorrect label from the AFRS but would

increase when the label was correct. To infer meaningful differences, these change values should exceed those that occurred when the images were shown in the control condition (i.e., without a decision label from the AFRS).

Method

Participants

The final sample consisted of 38 participants in the AFRS-High condition (M = 34.1, SD = 11.6, 27 females, 11 males), 47 in the AFRS-Low condition (M = 32.7, SD = 9.6, 35 females, 11 males, 1 other) and 45 in the control condition (M = 33.4, SD = 10.8, 32 females, 12 males, 1 other).

Design

Aid Condition (AFRS-High, AFRS-Low, control) was a between-participants factor, and Task Phase (baseline, test) was a within-participants factor. Table 5 shows the baseline human accuracy for the pairs selected to be errors for AFRS-High and AFRS-Low.

Table 5

Baseline Human Accuracy (n = 383; Set A = 192, Set B = 191) for the match and mismatch pairs selected to be Easy (AFRS-High) and Hard (AFRS-Low) errors in Experiment 3, shown separately for EFCT Sets A and B. All values show accuracy (%).

		Set A					Set B				
Errors	**Identity**	**Pair 1**	**Pair 2**	**Pair 3**	**Pair 4**	**Avg.**	**Pair 1**	**Pair 2**	**Pair 3**	**Pair 4**	**Avg.**
Easy	Match	96.4	96.9	97.9	99.0	**97.5**	89.0	91.1	93.7	95.3	**92.3**
	Mismatch	94.8	96.9	96.9	97.4	**96.5**	95.3	96.3	96.9	98.4	**96.7**
Hard	Match	21.9	31.8	34.4	35.9	**31.0**	22.5	34.6	36.6	42.4	**34.0**
	Mismatch	32.8	56.8	59.9	62.5	**53.0**	34.6	68.1	68.6	69.6	**60.2**

Results

Participant Data

Sensitivity

The main effects of Task Phase, $F(1, 127) = 55.06$, $p < .001$, $\eta_p^2 = .30$, and Aid Condition, $F(2, 127) = 3.74$, $p = .026$, $\eta_p^2 = .06$, were significant (see Figure 6a), as was their interaction, $F(2, 127) = 18.19$, $p < .001$, $\eta_p^2 = .22$. Sensitivity increased in the test phase for

both AFRS conditions, but not the control condition (see Table 6). The increase in sensitivity did not differ between the two AFRS conditions (see supplementary materials).

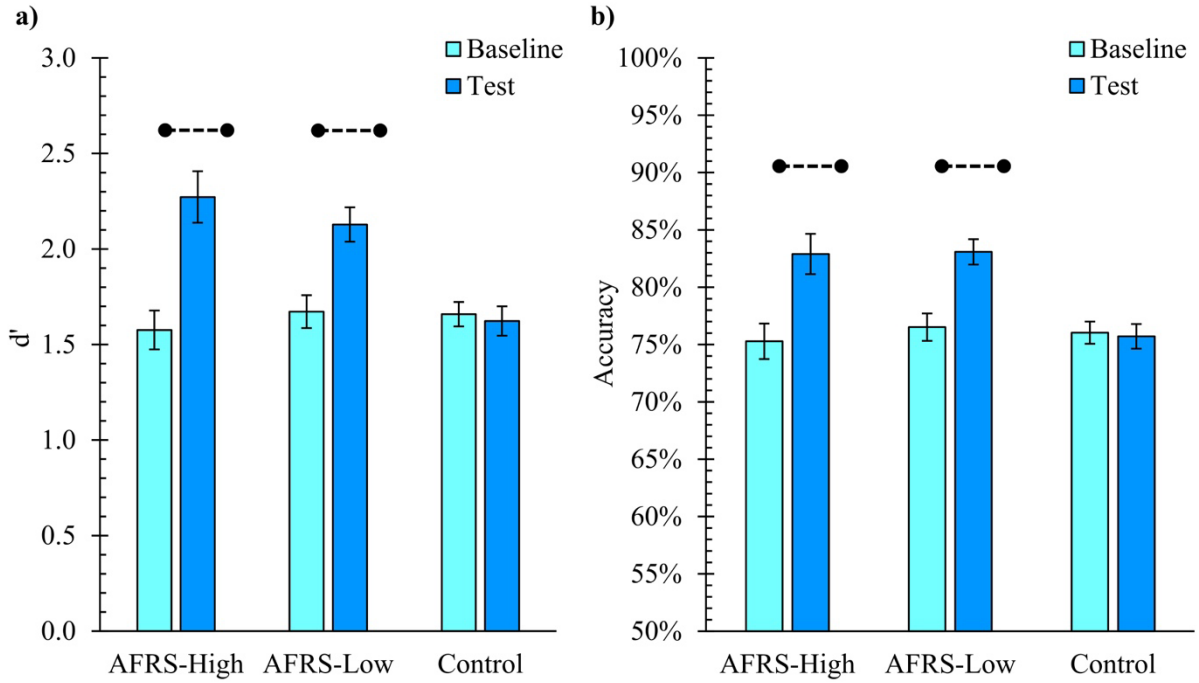

Figure 6. a) Sensitivity and **b)** Overall Accuracy for each Aid Condition (AFRS-High, AFRS-Low, control) in the Task Phases (baseline, test) of Experiment 3. The dashed bars show the performance of each AFRS alone. All error bars show the SEM.

In contrast to our prediction, the aided performance of the AFRS-High condition at test was significantly lower than that of their AFRS alone (d′ = 2.618), $t(37) = -2.57$, 95%CI[-0.62, -0.07], $p = .014$, $d = -0.42$, $BF_{10} = 3.06$. The aided test performance of the AFRS-Low condition was also below that of their AFRS (d′ = 2.618), $t(46) = -5.47$, 95%CI[-0.67, -0.31], $p < .001$, $d = -0.80$, $BF_{10} = 9{,}548$.

Table 6

Simple main effects analyses for Experiment 3, showing sensitivity and overall accuracy for each Aid Condition (AFRS-High, AFRS-Low, control). 95%CI are given for the difference between the Baseline and Test Phases.

Measure	Aid	Baseline	Test	*t*	*df*	95%CI	*p*	*d*	BF_{10}
Sensitivity	AFRS-H	1.58 (0.63)	2.27 (0.83)	5.94	37	0.46, 0.93	< .001*	0.96	22,323
	AFRS-L	1.67 (0.59)	2.13 (0.62)	5.65	46	0.29, 0.62	< .001*	0.83	17,297
	Control	1.66 (0.43)	1.62 (0.52)	-0.58	44	-0.16, 0.09	.567	0.09	0.189
Accuracy	AFRS-H	75.28 (9.55)	82.90 (10.81)	5.18	37	4.64, 10.59	< .001*	0.84	2,468
	AFRS-L	76.52 (8.18)	83.08 (7.62)	6.30	46	4.46, 8.66	< .001*	0.92	138,987
	Control	76.03 (6.46)	75.71 (7.23)	-0.34	44	-2.18, 1.54	.733	0.05	0.171

Overall Accuracy

The main effect of Task Phase was significant, $F(1, 127) = 49.39$, $p < .001$, $\eta_p^2 = .28$, as was the main effect of Aid Condition, $F(2, 127) = 3.56$, $p = .031$, $\eta_p^2 = .05$, and the interaction between the two factors, $F(2, 127) = 14.48$, $p < .001$, $\eta_p^2 = .19$ (see Figure 6b). Overall accuracy increased at test for the AFRS-High and AFRS-Low conditions, but not the control condition (see Table 6).

Criterion

The main effect of Task Phase was significant, $F(1, 127) = 29.08$, $p < .001$, $\eta_p^2 = .19$, with a larger conservative response bias at baseline ($M = 0.27$, $SD = 0.44$) than at test ($M = 0.12$, $SD = 0.46$). The main effect of Aid Condition was non-significant, $F(2, 127) = 1.54$, $p = .219$, $\eta_p^2 = .02$. The interaction was also non-significant, $F(2, 127) = 0.20$, $p = .819$, $\eta_p^2 = .00$.

Image Pair Analysis

Finally, we examined the effect of the AFRS decision label on the accuracy for each face pair[5]. By counterbalancing the presentation order of EFCT Sets A and B, each AFRS (-High, -Low) gave correct answers to 152 face pairs and made errors on 8 match and 8 mismatch trials across participants. But because AFRS-High made errors on the pairs with the highest average human accuracy ("easy errors"), while AFRS-Low erred on those with the lowest ("hard errors"), each error pair was also shown with the correct label to some participants. Thus, there are 136 face pairs that were only ever shown with the correct label, while the 16 error pairs from AFRS-High and 16 error pairs from AFRS-low were shown with correct and incorrect labels. Each pair was also shown without a decision label in the control condition.

[5] Figure 7 represents a minor deviation from our pre-registered analysis plan as we have not plotted the change in accuracy separately for images from Sets A and B of the EFCT. A very similar pattern of results occurred for images from Sets A and B, and these data are available on the OSF.

When the AFRS gave the correct decision, accuracy increased across all image conditions (see Figure 7). Conversely, an incorrect decision from the AFRS resulted in decreased accuracy across both error conditions. The change in accuracy (both positive and negative) for the "hard" errors was more extreme than for the "easy" errors, which is consistent with the suggestion that participants relied more on the AFRS for the most difficult pairs, potentially because they were less certain about their own answer to those trials. For both correct and incorrect AFRS decision labels, the change in accuracy exceeded that of the control condition (less than 3% in each condition), demonstrating that the AFRS decision labels had a significant effect on human responses. Similar patterns were observed when examining match and mismatch trials separately (see supplementary materials).

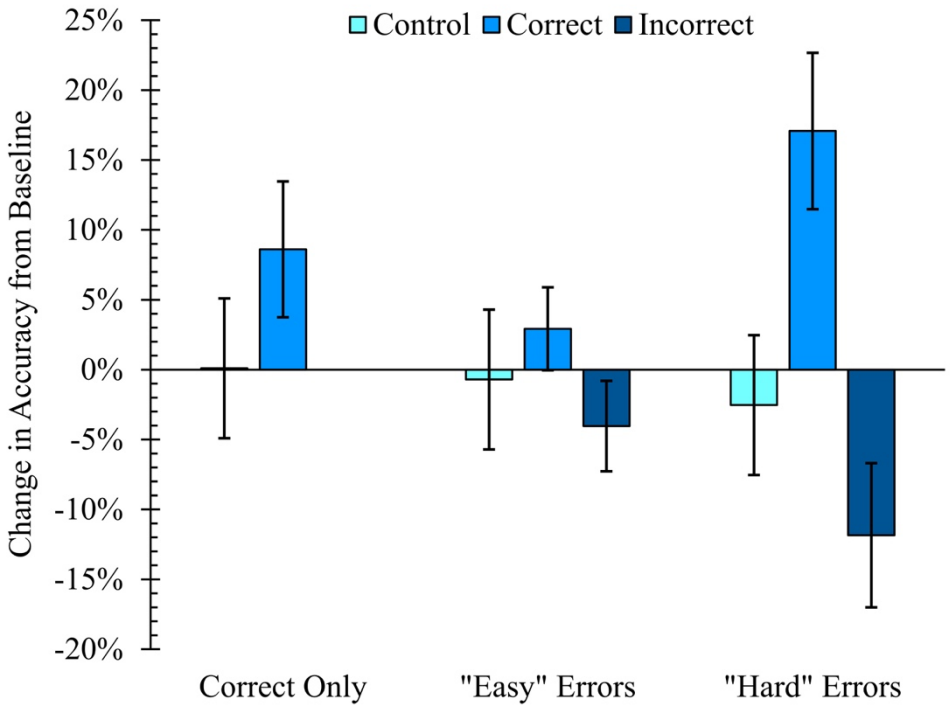

Figure 7. The average change in accuracy (test minus baseline) from baseline for face pairs that were always shown with the correct label (n = 136), those that were the "easy" errors (n = 16) for AFRS-High, and those that were "hard" errors (n = 16) for AFRS-Low, when shown with control, correct, and incorrect AFRS decision labels. Error bars show 1SD around the mean.

Discussion

Contrary to our predictions, aided performance at test did not differ between the AFRS-High and AFRS-Low conditions. Participants improved their sensitivity compared to baseline when aided by either AFRS. Although we expected that participants would achieve higher collaborative performance by overruling AFRS-High on obvious errors and following its other decisions, this did not occur. Instead, the aided test phase performance of the AFRS-High condition failed to reach that of the AFRS alone, as did that of the AFRS-Low condition. Notably, average item accuracy still fell when AFRS-High made errors, demonstrating that human operators are not certain to correct the AFRS, even when it makes errors that should be obvious to most observers (baseline accuracy > 89%). Taken together, these data raise questions about which conditions, if any at all, would enable human operators to produce AFRS-aided performance exceeding that of a highly accurate AFRS alone.

General Discussion

Summary

Across five pre-registered experiments, we have shown that human operators can improve their face matching performance when using a highly accurate AFRS as a decision-aid. However, aided human performance failed to reach, let alone exceed, the level of performance that each highly accurate AFRS offered alone. Supplementing the AFRS's binary identity decision with a similarity value did not further increase aided performance. But we also report some encouraging findings. Human performance was not impaired by the low accuracy AFRS55, indicating that participants did not follow the system's decisions uncritically. Moreover, this result was replicated when participants were unaware of the system's exact accuracy, suggesting that operators might be able to gauge the approximate accuracy of their AFRS over time. Ultimately, however, participants often overruled correct decisions from the AFRS and failed to correct errors, suggesting that they had little insight

into the accuracy of the AFRS on any one trial. Considered together, these findings offer further support for the notion that human ability likely limits the collaborative performance of a human-algorithm face verification team (White, Dunn, et al., 2015). These findings have implications for "human-in-the-loop" models of AFRS oversight.

Implications for Applied Settings

Many state-of-the-art AFRS can outperform most humans on face matching tasks (Phillips et al., 2018); for example, the DCNN used as the basis for the simulated AFRS in these experiments achieved 100% accuracy on the EFCT (see supplementary materials), while the average baseline accuracy of our participants was between 74-78% in each experiment. However, real AFRS can still make surprising errors (Hancock et al., 2020), and as such, ABC e-Gates require human oversight (FRONTEX, 2015; Fysh & Bindemann, 2018a; MacLeod & McLindin, 2011). This directive is presumably based on the assumption that the human operator will catch and overturn any incorrect decisions from the AFRS. But in Experiment 1, we found that an incorrect decision from the AFRS caused item accuracy to fall by an average of 5% for match trials and 15% for mismatches compared to baseline (or 10% and 12%, respectively, if compared to the change in performance of the control condition in the same task phase). Crucially, accuracy also fell in Experiment 3 when AFRS-High made errors on face pairs with the highest accuracy among human observers. These findings demonstrate that humans are not certain to overturn errors from the AFRS, even when the correct decision should be clear to most observers.

Sub-optimal Human-Computer Interaction

Several factors can help to explain why our participants often failed to correct errors from the AFRS. First, unfamiliar face matching is simply an error-prone task among both ordinary observers (Burton et al., 2010; Megreya & Burton, 2006), and many professional groups (White et al., 2014; Wirth & Carbon, 2017). As noted in the introduction, average

accuracy on many standardised face matching tasks falls in the range of 70-90% correct (Burton et al., 2010; Carragher & Hancock, 2020; Fysh & Bindemann, 2018b). Second, the benefit of human-algorithm teaming would be greatest if the human and AFRS were uncorrelated in their errors (e.g., if humans were just as likely to correctly resolve a face pair that the AFRS errs on as they are any randomly selected pair). But Experiment 3 revealed significant correlations between human accuracy and the real DCNN's similarity ratings for match and mismatch pairs. Thus, not only is unfamiliar face matching already a difficult task (Jenkins et al., 2011), but it is likely even harder for the face pairs that AFRS tend to err on. Concerningly, it is possible that these issues might be compounded by ethnicity, both of the human operator and the faces being examined. With reports of racial biases in some AFRS (Grother et al., 2019; Phillips, Jiang, et al., 2011), and other race effects (Meissner & Brigham, 2001) documented in the unfamiliar face matching performance of humans (Megreya et al., 2011; Meissner et al., 2013), further research is needed to examine how human-AFRS teams perform when matching faces of various ethnicities.

We must also consider how human operators tend to use automated decision-aids (Parasuraman & Riley, 1997). Despite improving compared to baseline, aided human performance failed to surpass that of the AFRS alone because participants often overruled the system's correct decisions. This finding is consistent with previous reports of sub-optimal performance in human-automation teams (Bartlett & McCarley, 2017; Boskemper et al., 2021). An operator's reliance on their decision-aid is strongly influenced by their trust in the aid and the confidence they have in their own ability (Lee & Moray, 1994; Lee & See, 2004). Automation use is more likely when trust in the system is high and self-confidence is low, whereas disuse can occur when trust in the system is low and self-confidence is high (Lee & Moray, 1994). Because humans only have moderate insight into their own general face identification abilities (Bobak et al., 2019; Zhou & Jenkins, 2020), and may know little about

the capabilities of modern AFRS (Ritchie et al., 2021), naïve participants might struggle to weigh the AFRS decision against their own appropriately (Hoff & Bashir, 2015), which is required for optimal collaborative decision making (Bahrami et al., 2010; Sorkin et al., 2001). Finally, it should be considered that the propensity of our participants to overrule the AFRS, even when it was correct, suggests that the collaborative performance of human-AFRS teams is not likely to be improved just by focusing on increasing the accuracy of the AFRS. Further research is needed to examine strategies to improve the ability of the human to weigh the decision from the AFRS appropriately.

Improving Human-AFRS Oversight

If the performance of human-algorithm teams is limited by human ability, how can necessary human oversight be applied to enhance the performance of an AFRS? In the introduction, we discussed "fusion" (O'Toole et al., 2007), a process wherein combining independent judgments from humans and AFRS in a weighted average can produce gains in accuracy above either individual source (Phillips et al., 2018). Although this procedure does not reflect the sequential nature of some oversight models (Fysh & Bindemann, 2018a), fusion might offer a route to improved human-AFRS teaming in operational settings. By asking the human operator to make their judgment independently of the AFRS, complex issues relating to the AFRS decision biasing the human response can be avoided (Howard et al., 2020). Moreover, a fusion algorithm could potentially be used to weigh the independent judgments of the human operator and the AFRS by the accuracy of their past performance in similar circumstances, a task left to the human operator in the current paradigm. Future research should investigate whether a fusion-based approach might produce the collaborative performance gains that were lacking from our simplified sequential model of human-in-the-loop AFRS oversight. However, any such approach would still need to provide a mechanism for the human operator to flag any egregious errors from the AFRS (Hancock et al., 2020).

Limitations & Future Directions

This project was designed to answer several basic questions about the collaborative performance of human-AFRS teams on tasks of one-to-one face matching. However, there are limitations to be acknowledged and addressed in future studies. First, because our participants were lay people who were recruited in exchange for a small payment, it is possible that they did not undertake the task with the seriousness expected of professionals who use these systems daily. Moreover, experience using an accurate automated aid can contribute to trust in the system (Hoff & Bashir, 2015), which in turn can lead to greater reliance on the decision-aid (Lee & Moray, 1994; Lee & See, 2004). As such, future research should investigate whether individuals who use AFRS in their work (e.g., border control officers) are able to achieve levels of AFRS-aided performance higher than those reported in the current study. But the fact that our sample consisted of lay people does not invalidate our conclusions. Many professionals have average face identification abilities (Weatherford et al., 2021; White et al., 2014), even when they are incentivised to perform well (Kemp et al., 1997). This work also speaks to the issues that may arise with placing lay people in roles that require evaluation of AFRS decisions, which will likely become more common with the proliferation of AFRS technologies across sectors (Centre for Data Ethics and Innovation, 2020; Noyes & Hill, 2021; Ritchie et al., 2021).

Second, although we advanced on previous methodologies by basing the performance of our simulated AFRS on that of a real DCNN (Carragher & Hancock, 2020), the real DCNN did not make any errors on the EFCT (see supplementary materials). As such, we selected "error" pairs based on the proximity of their similarity rating to the DCNN's decision threshold. Of course, the realism of future research would benefit from using an AFRS that makes genuine errors on the chosen stimulus set. Yet, there are also reasons that this approach was not optimal here. For example, the images in the EFCT are relatively high

quality, showing the subject front on and free of occlusions (White, Phillips, et al., 2015). Many state-of-the-art AFRS perform with remarkable accuracy under these conditions (Grother et al., 2021), as should those that are used in "e-Gates" (FRONTEX, 2015). Nonetheless, future research is needed to replicate our results with different AFRS, because the performance of every system on the EFCT will be unique (either in accuracy or similarity ratings). Alternatively, future researchers might choose to investigate the collaborative performance of human-AFRS teams for highly challenging tasks with low quality images that are likely to create genuine performance errors.

Finally, it should be considered that we have presented a simplified model of human oversight of AFRS decisions in identity verification tasks. Our paradigm was based upon those implemented by Fysh and Bindemann (2018a) and Howard et al. (2020), along with documented accounts of workflows used in some ABC e-Gates scenarios (FRONTEX, 2015; MacLeod & McLindin, 2011). However, it is entirely possible, if not likely, that there will be operational differences in how human oversight is implemented by different organisations or in different applied settings. Yet, our aim was not to test a single specific model of human AFRS oversight. Rather, we sought to investigate how knowing the prior decision of the AFRS would affect the final identification decision offered by the human operator. Our central finding, that humans overrule correct AFRS decisions while also failing to correct errors, is relevant to the implementation, whether existing or planned, of human AFRS oversight in many scenarios.

Conclusion

Automated facial recognition systems are becoming more common in society, whether securing sensitive infrastructure or in public surveillance and national security settings (Noyes & Hill, 2021). Despite significant advances in accuracy, human oversight of these systems is still required to catch errors or resolve inconclusive judgments (Fysh &

Bindemann, 2018a; White, Dunn, et al., 2015). Using a simulated AFRS and samples of lay participants, we demonstrate that despite significant increases to human accuracy, the AFRS-aided performance of the human operator fails to equal that of the AFRS alone. Concerningly, participants did not always correct the system when it made errors that should be obvious to most humans, which is one of the central reasons why humans are used to perform oversight of AFRS decisions. These data strongly suggest that, at least in the current paradigm, human performance is likely limiting the potential benefits of AFRS, both by failing to correct errors and by overruling correct decisions. Careful consideration must be given as to how the necessary human oversight of AFRS can be implemented in a way that enhances the performance of the system.

Declarations

Ethics approval and consent to participate

All participants gave their informed consent before starting the experiment. This research was approved by the General University Ethics Panel at the University of Stirling [#GUEP502].

Consent for publication

Figure 1 of this manuscript has been published in accordance with the terms of the license governing the use of the EFCT.

Availability of data and materials

All datasets analysed in the current study and supplementary materials are available in the OSF repository [https://osf.io/d4vkm/].

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by an Engineering and Physical Sciences Research Council grant to PJBH (#EP/N007743/1). The funding body had no role in this study.

Author Contributions

Both authors contributed to the conception of this project and helped design each experiment. DJC programmed the experiments and oversaw human data collection. PJBH oversaw data collection from the DCNN. DJC and PJBH analysed and interpreted the data. DJC wrote the manuscript. PJBH provided critical revisions to the manuscript.

Acknowledgements

The authors wish to thank Professor Jason McCarley and Dr Megan Bartlett for some very helpful conversations early in the life of this project. DJC would also like to thank Pete and Ali Richardson – I ran these five experiments from your kitchen table while waiting to return to Australia. Your incredible kindness will never be forgotten.

Open Practices Statement

Prior to data collection, we pre-registered the aims, hypotheses, design, and analyses for the current study on the OSF. The datasets generated and analysed in the main text and supplementary materials are also available in the same OSF repository [https://osf.io/d4vkm/].

References

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. https://doi.org/10.7717/peerj.1184

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-1085. https://doi.org/10.1126/science.1185718

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human factors*, *59*(6), 881-900. https://doi.org/10.1177/0018720817700258

Bartlett, M. L., & McCarley, J. S. (2019). No effect of cue format on automation dependence in an aided signal detection task. *Human factors*, *61*(2), 169-190. https://doi.org/10.1177/0018720818802961

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLoS One*, *11*(2), e0148148. https://doi.org/10.1371/journal.pone.0148148

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. https://doi.org/10.1177/1747021818776145

Boskemper, M. M., Bartlett, M. L., & McCarley, J. S. (2021). Measuring the Efficiency of Automation-Aided Performance in a Simulated Baggage Screening Task. *Human factors*, 0018720820983632. https://doi.org/10.1177/0018720820983632

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339-360. https://doi.org/10.1037/1076-898x.5.4.339

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286-291. https://doi.org/10.3758/brm.42.1.286

Carragher, D., Towler, A., Mileva, V. R., White, D., & Hancock, P. J. (2022). Masked face identification is improved by diagnostic feature training. *Cognitive Research: Principles and Implications, 7(1), 1-12*. https://doi.org/10.1186/s41235-022-00381-x

Carragher, D. J., & Hancock, P. J. (2020). Surgical face masks impair human face matching performance for familiar and unfamiliar faces. *Cognitive Research: Principles and Implications*, *5*(1), 1-15. https://doi.org/10.1186/s41235-020-00258-x

Centre for Data Ethics and Innovation. (2020). *Facial recognition technology*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/905267/Facial_Recognition_Technology_Snapshot_UPDATED.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/bf03193146

FRONTEX. (2015). *Best Practice Technical Guidelines for Automated Border Control (ABC) Systems*. European Agency for the Management of Operational Cooperation at the … Retrieved from https://frontex.europa.eu/assets/Publications/Research/Best_Practice_Technical_Guidelines_ABC.pdf

REFERENCES

Fysh, M. C., & Bindemann, M. (2018a). Human–Computer Interaction in Face Matching. *Cognitive science*, *42*(5), 1714-1732. https://doi.org/10.1111/cogs.12633

Fysh, M. C., & Bindemann, M. (2018b). The Kent face matching test. *British Journal of Psychology*, *109*(2), 219-231. https://doi.org/10.1111/bjop.12260

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, *75*(7), 450-451.

Goss-Sampson, M., van Doorn, J., & Wagenmakers, E. (2020). Bayesian inference in JASP: A guide for students. *University of Amsterdam: JASP team*.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology. https://doi.org/10.6028/nist.ir.8280

Grother, P., Ngan, M., Hanaoka, K., Yang, J. C., & Hom, A. (2021). *Ongoing Face Recognition Vendor Test (FRVT). Part 1: Verification*. Retrieved from https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330-337. https://doi.org/10.1016/s1364-6613(00)01519-9

Hancock, P. J., Somai, R. S., & Mileva, V. R. (2020). Convolutional neural net face recognition works in non-human-like ways. *Royal Society Open Science*, *7*, 200595. https://doi.org/10.1098/rsos.200595

Heyer, R., Semmler, C., & Hendrickson, A. T. (2018). Humans and algorithms for facial recognition: The effects of candidate list length and experience on performance. *Journal of applied research in memory and cognition*, *7*(4), 597-609. https://doi.org/10.1016/j.jarmac.2018.06.002

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

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), e0237855. https://doi.org/10.1371/journal.pone.0237855

JASP Team. (2020). JASP (Version 0.14.0)[Computer software].

Jeckeln, G., Hahn, C. A., Noyes, E., Cavazos, J. G., & O'Toole, A. J. (2018). Wisdom of the social versus non-social crowd in face identification. *British Journal of Psychology*, *109*(4), 724-735. https://doi.org/10.1111/bjop.12291

Jenkins, R., White, D., Van Montfort, X., & Burton, M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323. https://doi.org/10.1016/j.cognition.2011.08.001

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*(3), 211-222. https://doi.org/10.1002/(sici)1099-0720(199706)11:3<211::aid-acp430>3.0.co;2-o

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153-184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50.30392

MacLeod, V., & McLindin, B. (2011). Methodology for the evaluation of an international airport automated border control processing system. In *Innovations in Defence Support Systems-2* (pp. 115-145). Springer.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Matsuyoshi, D., & Watanabe, K. (2021). People have modest, not good, insight into their face recognition ability: a comparison between self-report questionnaires. *Psychological Research*, *85*(4), 1713-1723. https://doi.org/10.1007/s00426-020-01355-8

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876. https://doi.org/10.3758/bf03193433

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*(7), 1175-1184. https://doi.org/10.3758/bf03193954

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology*, *64*(8), 1473-1483. https://doi.org/10.1080/17470218.2011.575228

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3-35. https://doi.org/10.1037//1076-8971.7.1.3

Meissner, C. A., Susa, K. J., & Ross, A. B. (2013). Can I see your passport please? Perceptual discrimination of own-and other-race faces. *Visual Cognition*, *21*(9-10), 1287-1305. https://doi.org/10.1080/13506285.2013.832451

Ngan, M. L., Grother, P. J., & Hanaoka, K. K. (2020). *Ongoing Face Recognition Vendor Test (FRVT) Part 6A: Face recognition accuracy with masks using pre-COVID-19 algorithms*. (NISTIR8311).

Noyes, E., & Hill, M. Q. (2021). Automatic Recognition Systems and Human Computer Interaction in Face Matching. In *Forensic Face Matching: Research and Practice* (pp. 193-215). Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0009

O'Toole, A. J., Abdi, H., Jiang, F., & Phillips, P. J. (2007). Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *37*(5), 1149-1155. https://doi.org/10.1109/tsmcb.2007.907034

O'Toole, A. J., An, X., Dunlop, J., Natu, V., & Phillips, P. J. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception (TAP)*, *9*(4), 1-13. https://doi.org/10.1145/2355598.2355599

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D. S., Dunlop, J., Lui, Y. M., Sahibzada, H., & Weimer, S. (2011). An introduction to the good, the bad, & the ugly face recognition challenge problem. 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG),

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, *8*(2), 1-11. https://doi.org/10.1145/1870076.1870082

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

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., & Sankaranarayanan, S. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171-6176. https://doi.org/10.1073/pnas.1721355115

Ritchie, K. L., Cartledge, C., Growns, B., Yan, A., Wang, Y., Guo, K., Kramer, R. S., Edmond, G., Martire, K. A., & San Roque, M. (2021). Public attitudes towards the

use of automatic facial recognition technology in criminal justice systems around the world. *PLoS One*, *16*(10), e0258241. https://doi.org/10.1371/journal.pone.0258241

Robinson, D. E., & Sorkin, R. D. (1985). A contingent criterion model of computer assisted detection. *Trends in ergonomics/human factors*, *2*, 75-82.

Ross, J. M., Szalma, J. L., Hancock, P. A., Barnett, J. S., & Taylor, G. (2008). The effect of automation reliability on user automation trust and reliance in a search-and-rescue scenario. Proceedings of the Human Factors and Ergonomics Society Annual Meeting,

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183. https://doi.org/10.1037/0033-295x.108.1.183

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137-149. https://doi.org/10.3758/bf03207704

Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence–accuracy relationship in unfamiliar face matching. *Journal of Experimental Psychology: Applied*, *23*(3), 336. https://doi.org/10.1037/xap0000130

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

Weiss, B. A. (2011). *Fisher's r-to-Z transformation calculator to compare two independent samples*. In https://blogs.gwu.edu/weissba/teaching/calculators/fishers-z-transformation/

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology*, *27*(6), 769-777. https://doi.org/10.1002/acp.2971

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS One*, *10*(10), e0139827. https://doi.org/10.1371/journal.pone.0139827

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS One*, *9*(8), e103510. https://doi.org/10.1371/journal.pone.0103510

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. https://doi.org/10.1098/rspb.2015.1292

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201-212. https://doi.org/10.1080/14639220500370105

Wirth, B. E., & Carbon, C.-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

Zhou, X., & Jenkins, R. (2020). Dunning–Kruger effects in face perception. *Cognition*, *203*, 104345. https://doi.org/10.1016/j.cognition.2020.104345

Simulated Automated Facial Recognition Systems as Decision-Aids in Forensic Face Matching Tasks

*Daniel J. Carragher[1, 2] & Peter J. B. Hancock[1]

SUPPLEMENTARY MATERIALS

[1]Psychology

Faculty of Natural Sciences

University of Stirling

Scotland, United Kingdom

[2]School of Psychology

Faculty of Health and Medical Sciences

University of Adelaide

Adelaide, Australia

Total Word Count: 2,500 approx.

*Corresponding Author:
Daniel J. Carragher
School of Psychology
Faculty of Health and Medical Sciences
University of Adelaide
Adelaide, South Australia, 5000
daniel.carragher@adelaide.edu.au

Performance of the Deep Convolutional Neural Network

The real FACER2VM Deep Convolutional Neural Network (DCNN) correctly classified each face pair in the Expertise in Facial Comparison Test (EFCT). The datafile is available on the Open Science Framework (https://osf.io/d4vkm/). In the introduction to Experiment 3, we report the correlation between human accuracy on the EFCT (using baseline data from Experiment 1a, 2a and 2b, n = 383) and the similarity ratings from the DCNN. These data are also available in the same file on the OSF ("OSF_DCNN_EFCT_PerformanceData.xlsx").

Experiment 1a

Excluded Participants

We initially received responses from 111 participants. Participants that finished the task too quickly (n = 1) or too slowly (n = 1) were excluded, as were those that did not complete the experiment (n = 1) or attempted the face matching task more than once (n = 1). Finally, participants who gave an incorrect response to a question about their experimental condition were also excluded (n = 8). In total, 2 participants were excluded from the AFRS condition, and 10 from the control condition.

Results

Accuracy

Match Trials. The main effect of Task Phase was significant, $F(1, 97) = 33.91$, $p < .001$, $\eta_p^2 = .26$, as was the main effect of Aid Condition, $F(1, 97) = 5.78$, $p = .018$, $\eta_p^2 = .06$, and their interaction, $F(1, 97) = 8.59$, $p = .004$, $\eta_p^2 = .08$ (see Figure S1). There was greater improvement in match trial accuracy at test among the AFRS condition ($F = 39.57$, $p < .001$), than the control condition ($F = 4.11$, $p = .049$).

Mismatch Trials. The main effect of Task Phase was non-significant, $F(1, 97) = 0.01$, $p = .925$, $\eta_p^2 = .00$, as was the main effect of Aid Condition, $F(1, 97) = 0.58$, $p = .448$, η_p^2

= .01. However, the interaction between the two factors was significant, $F(1, 97) = 9.17$, $p = .003$, $\eta_p^2 = .09$ (see Figure S1), due to the significant increase in mismatch accuracy for the AFRS condition at test ($F = 4.28$, $p = .043$), and significant decrease for the control condition ($F = 5.03$, $p = .030$).

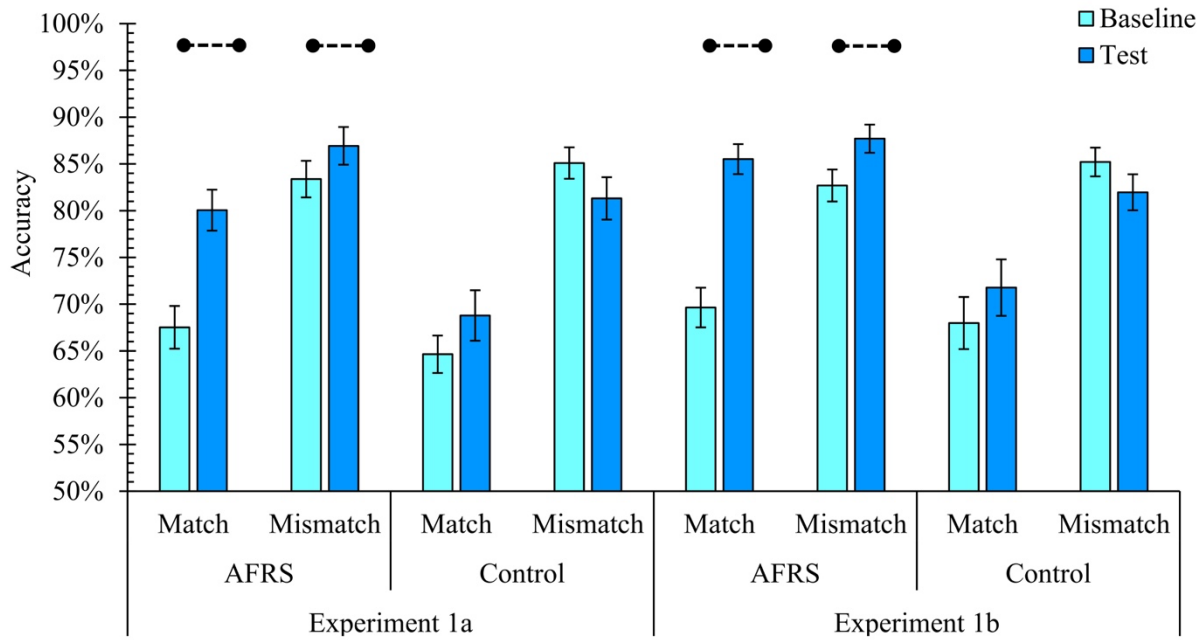

Figure S1. Accuracy for match and mismatch pairs in the baseline and test task phases, plotted separately for each Aid Condition (AFRS, control) in Experiment 1a and Experiment 1b. The dashed bars show the performance of the AFRS alone. All error bars show the standard error of the mean (SEM).

Criterion

The conservative response bias shown by participants in both aid conditions differed significantly from 0 (no bias) in the baseline and test task phases (see Table S1).

Table S1

One sample *t*-tests comparing the response bias shown by each Aid Condition (AFRS, control) in Experiment 1 (a, b) to 0, in order to determine whether the response bias is statistically significant.

Experiment	Aid	Task Phase	*M (SD)*	*df*	*t*	95%CI	*p*	*d*
1a	AFRS	Baseline	0.33 (0.50)	52	4.79	0.19, 0.46	< .001*	0.66
		Test	0.17 (0.47)	52	2.59	0.04, 0.30	.012*	0.36
	Control	Baseline	0.40 (0.36)	45	7.51	0.29, 0.50	< .001*	1.11
		Test	0.24 (0.54)	45	3.07	0.08, 0.40	.004*	0.45
1b	AFRS	Baseline	0.25 (0.44)	47	3.96	0.12, 0.38	< .001*	0.57
		Test	0.06 (0.41)	47	1.04	-0.06, 0.18	.302	0.15
	Control	Baseline	0.34 (0.50)	46	4.65	0.19, 0.48	< .001*	0.68
		Test	0.21 (0.60)	46	2.35	0.03, 0.38	.023*	0.34

*Identifies statistically significant comparisons.

Experiment 1b

Excluded Participants

We initially received responses from 113 participants. Participants that finished the task too quickly (n = 4) or too slowly (n = 2) were excluded, as were those who did not complete the experiment (n = 3) or attempted the face matching task more than once (n = 1). Finally, participants who failed an attention check trial (n = 2) or who gave an incorrect response to a question about their experimental condition were also excluded (n = 6). In total, 7 participants were excluded from the AFRS condition, and 10 from the control condition. One respondent ended their participation before being assigned a condition.

Results

Accuracy

Match Trials. The main effect of Task Phase was significant, $F(1, 93) = 63.99$, $p < .001$, $\eta_p^2 = .41$, as was the main effect of Aid Condition, $F(1, 93) = 5.72$, $p = .019$, $\eta_p^2 = .06$, and their interaction, $F(1, 93) = 24.11$, $p < .001$, $\eta_p^2 = .21$ (see Figure S1). There was a greater improvement in match trial accuracy at test among the AFRS condition ($F = 89.31$, $p < .001$), than the control condition ($F = 4.46$, $p = .040$).

Mismatch Trials. The main effect of Task Phase was non-significant, $F(1, 93) = 0.82$, $p = .369$, $\eta_p^2 = .01$, as was the main effect of Aid Condition, $F(1, 93) = 0.55$, $p = .459$, $\eta_p^2 = .01$. However, the interaction between the two factors was significant, $F(1, 93) = 17.79$, $p < .001$, $\eta_p^2 = .16$ (see Figure S1), due to the significant increase in accuracy for the AFRS condition at test ($F = 12.51$, $p < .001$), and significant decrease for the control condition ($F = 5.79$, $p = .020$).

Criterion

The conservative response bias shown by participants in both aid conditions differed significantly from 0 at baseline (see Table S1). This bias remained significant in the test task phase for participants in the control condition, but not the AFRS condition.

Further Analysis: is Baseline Sensitivity related to Effective AFRS use?

In Experiment 1a, the relationship between an individual's baseline sensitivity and the change to their performance in the test task phase (test d′ minus baseline d′) was non-significant. However, re-testing this relationship in Experiment 1b revealed a significant relationship, $r(47) = -0.31$, 95%CI[-0.55, -0.03], $p = .032$. The differing results suggest that each correlation is likely underpowered at these sample sizes (only using participants from the AFRS condition).

To address this issue, we combined the data from Experiments 1a and 1b, since there was no significant effect of "decision-type" below. The relationship again proved to be non-significant, $r(100) = -0.19$, 95%CI[-0.37, -0.01], $p = .060$, $BF_{10} = 0.712$. It is unlikely that an individual's face matching ability offers significant insight into their ability to use the AFRS effectively. This conclusion supports that offered in the main text of Experiment 1a.

Comparison of Experiments 1a and 1b: The Effect of Decision-Type

Here we report the full results of the mixed measures ANOVA with Task Phase (baseline, test) as a within-participants factor, and Aid Condition (AFRS, control) and

Decision-Type (binary, similarity) as between-participants factors. The main effect of Decision-Type was non-significant, $F(1, 190) = 0.97$, $p = .326$, $\eta_p^2 = .01$. The main effects of Task Phase, $F(1, 190) = 86.34$, $p < .001$, $\eta_p^2 = .31$, and Aid Condition were both significant, $F(1, 190) = 19.25$, $p < .001$, $\eta_p^2 = .09$. The two way interactions between Task Phase and Decision-Type, $F(1, 190) = 1.67$, $p = .197$, $\eta_p^2 = .01$, and Aid Condition and Decision-Type were non-significant, $F(1, 190) = 0.01$, $p = .905$, $\eta_p^2 = .00$. As in the main text, the interaction between Task Phase and Aid Condition was significant, $F(1, 190) = 65.96$, $p < .001$, $\eta_p^2 = .26$, due to the significant increase in sensitivity at test among the AFRS condition ($F = 118.66$, $p < .001$), but not the control condition ($F = 1.03$, $p = .313$). The three-way interaction between the factors was non-significant, $F(1, 190) = 1.29$, $p = .258$, $\eta_p^2 = .01$.

Experiment 2a

Excluded Participants

We initially received responses from 168 participants. Participants that finished the task too quickly ($n = 6$) or too slowly ($n = 1$) were excluded, as were those who did not complete the experiment ($n = 3$) or attempted the face matching task more than once ($n = 2$). Finally, participants who failed an attention check trial ($n = 1$) or who gave an incorrect response to a question about their experimental condition were also excluded ($n = 19$). In total, 10 participants were excluded from the AFRS93 condition, 11 from the AFRS55 condition, and 10 from the control condition. One respondent ended their participation before being assigned a condition.

Results

Accuracy

Match Trials. The main effect of Task Phase was significant, $F(1, 133) = 28.10$, $p < .001$, $\eta_p^2 = .17$, as was the main effect of Aid Condition, $F(2, 133) = 4.78$, $p = .010$, $\eta_p^2 = .07$,

and their interaction, $F(2, 133) = 5.15$, $p = .007$, $\eta_p^2 = .07$ (see Figure S2). Match trial accuracy improved at test for the AFRS93 ($F = 26.75$, $p < .001$) and control conditions ($F = 5.91$, $p = .019$), but not the AFRS55 condition ($F = 1.87$, $p = .179$).

Mismatch Trials. The main effect of Task Phase was significant, $F(1, 133) = 4.75$, $p = .031$, $\eta_p^2 = .03$, as was the main effect of Aid Condition, $F(2, 133) = 5.87$, $p = .004$, $\eta_p^2 = .08$. The interaction between the two factors was also significant, $F(2,133) = 11.80$, $p < .001$, $\eta_p^2 = .15$ (see Figure S2), due to the significant increase in accuracy at test for the AFRS93 condition ($F = 5.30$, $p = .026$), and significant decrease in accuracy for the AFRS55 ($F = 7.33$, $p = .010$) and control conditions ($F = 15.33$, $p < .001$).

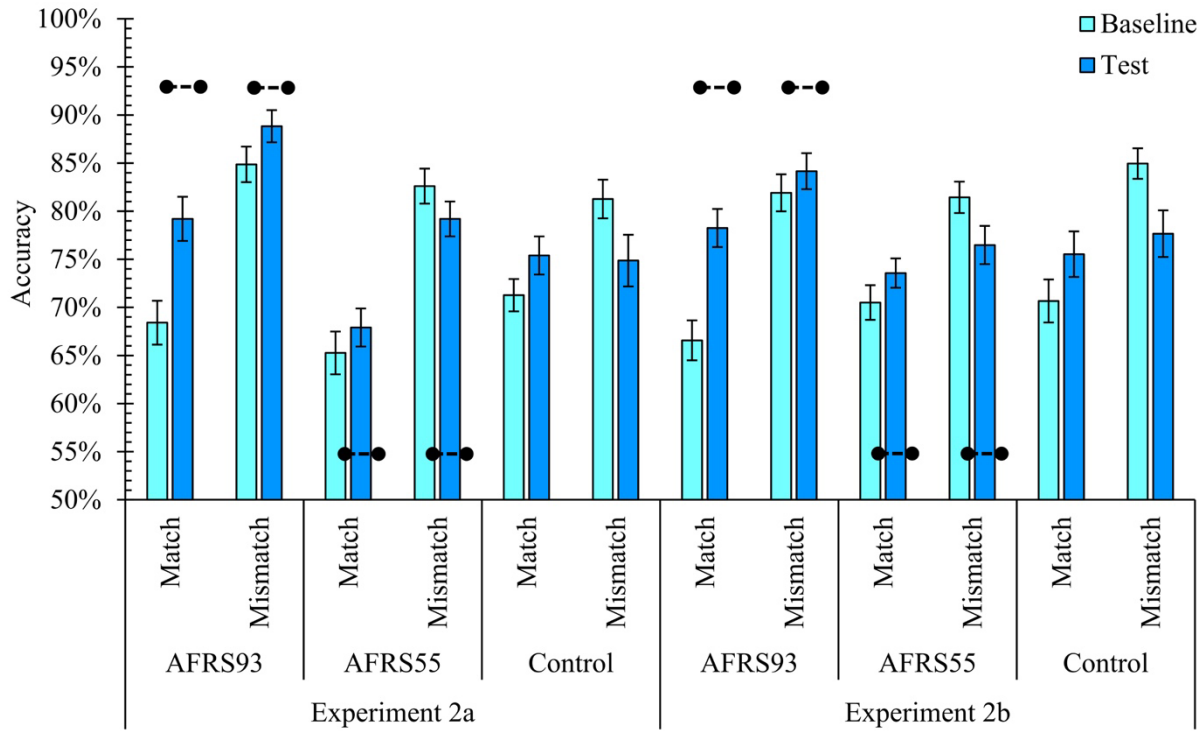

Figure S2. Accuracy for match and mismatch pairs in the baseline and test task phases, plotted separately for each Aid Condition (AFRS93, AFRS55, control) in Experiment 2a and Experiment 2b. The dashed bars show the performance of the AFRS alone. All error bars show the SEM.

Criterion

The conservative response bias was significant for all three aid conditions at baseline (see Table S2). This bias remained significant in the test task phase for the AFRS93 and AFRS55 conditions, but not the control condition.

Table S2

One sample *t*-tests comparing the response bias shown by each Aid Condition (AFRS93, AFRS55, control) in Experiment 2 (a, b) to 0, in order to determine whether the response bias is statistically significant.

Experiment	Aid	Task Phase	*M (SD)*	*df*	*t*	95%CI	*p*	*d*
2a	AFRS93	Baseline	0.33 (0.46)	44	4.73	0.19, 0.46	< .001*	0.71
		Test	0.25 (0.42)	44	3.90	0.12, 0.37	< .001*	0.58
	AFRS55	Baseline	0.31 (0.43)	45	4.97	0.19, 0.44	< .001*	0.73
		Test	0.20 (0.37)	45	3.69	0.09, 0.31	< .001*	0.54
	Control	Baseline	0.22 (0.39)	44	3.71	0.10, 0.33	< .001*	0.55
		Test	0.01 (0.46)	44	0.21	-0.13, 0.15	.838	0.03
2b	AFRS93	Baseline	0.30 (0.49)	51	4.39	0.16, 0.44	< .001*	0.61
		Test	0.13 (0.48)	51	1.92	-0.01, 0.26	.061	0.27
	AFRS55	Baseline	0.20 (0.30)	48	4.57	0.11, 0.28	< .001*	0.65
		Test	0.07 (0.32)	48	1.56	-0.02, 0.16	.125	0.22
	Control	Baseline	0.29 (0.42)	46	4.65	0.16, 0.41	< .001*	0.68
		Test	0.05 (0.49)	46	0.65	-0.10, 0.19	.518	0.10

*Identifies statistically significant comparisons.

Experiment 2b

Excluded Participants

We initially received responses from 167 participants. Participants that finished the task too quickly (n = 10) were excluded, as were those who did not complete the experiment (n = 3) or attempted the face matching task more than once (n = 2). Finally, participants who gave an incorrect response to a question about their experimental condition were also excluded (n = 4). In total, 3 participants were excluded from the AFRS93 condition, 6 from the AFRS55 condition, and 8 from the control condition. Two respondents ended their participation before being assigned a condition.

Results

Accuracy

Match Trials. The main effect of Task Phase was significant, $F(1, 145) = 35.54$, $p < .001$, $\eta_p^2 = .20$, but the main effect of Aid Condition was not, $F(2, 145) = 0.90$, $p = .914$, $\eta_p^2 = .00$. The interaction between the two factors was significant, $F(2, 145) = 5.90$, $p = .003$, $\eta_p^2 = .08$ (see Figure S2). Match trial accuracy improved at test for the AFRS93 ($F = 33.21$, $p < .001$) and control conditions ($F = 5.78$, $p = .020$), but not the AFRS55 condition ($F = 3.68$, $p = .061$).

Mismatch Trials. The main effect of Task Phase was significant, $F(1, 145) = 12.18$, $p < .001$, $\eta_p^2 = .08$, but the main effect of Aid Condition was not, $F(2, 145) = 1.42$, $p = .245$, $\eta_p^2 = .02$. The interaction between the two factors was significant, $F(2, 145) = 9.18$, $p < .001$, $\eta_p^2 = .11$ (see Figure S2). In contrast to Experiment 2a, there was no change to accuracy at test for the AFRS93 condition ($F = 2.18$, $p = .146$), but still a significant decrease in accuracy for the AFRS55 ($F = 10.89$, $p = .002$) and control conditions ($F = 14.15$, $p < .001$).

Criterion

Each aid condition showed a significant conservative response bias at baseline, which was non-significant at test (see Table S2).

Comparison between Experiment 2a and 2b: The Effect of Accuracy Foreknowledge

Here we report the full results of the mixed measures ANOVA with Task Phase (baseline, test) as a within-participants factor, and Aid Condition (AFRS93, AFRS55, control) and Accuracy Knowledge (known, unknown) as between-participants factors. The main effect of Accuracy Knowledge was not significant, $F(1, 278) = 0.01$, $p = .924$, $\eta_p^2 = .00$. The main effects of Task Phase, $F(1, 278) = 14.44$, $p < .001$, $\eta_p^2 = .05$, and Aid Condition were significant, $F(2, 278) = 14.73$, $p < .001$, $\eta_p^2 = .10$. The interactions between Accuracy

Knowledge and Task Phase, $F(1, 278) = 0.23$, $p = .629$, $\eta_p^2 = .00$, and Accuracy Knowledge and Aid Condition were both non-significant, $F(2, 278) = 2.36$, $p = .097$, $\eta_p^2 = .02$. The interaction between Task Phase and Aid Condition was significant, $F(2, 278) = 41.76$, $p < .001$, $\eta_p^2 = .23$, due to the significant increase in sensitivity for the AFRS93 condition at test ($F = 69.85$, $p < .001$). Sensitivity did not change for the AFRS55 condition ($F = 2.12$, $p = .149$), but neared a significant decrease for the control condition ($F = 3.93$, $p = .051$). The three-way interaction between the factors was non-significant, $F(2, 278) = 0.03$, $p = .972$, $\eta_p^2 = .00$.

Experiment 3

Excluded Participants

We initially received responses from 170 participants. Participants that finished the task too quickly ($n = 2$) or too slowly ($n = 4$) were excluded, as were those who did not complete the experiment ($n = 6$) or attempted the face matching task more than once ($n = 4$). Finally, participants who failed an attention check trial ($n = 3$) or gave an incorrect response to a question about their experimental condition were also excluded ($n = 21$). In total, 18 participants were excluded from the AFRS-High condition, 8 from the AFRS-Low condition, and 10 from the control condition. Four respondents ended their participation before being assigned a condition.

Results

Sensitivity

A one-way ANOVA on the change in d′ (test d′ minus baseline d′), $F(2, 77.14) = 20.68$, $p < .001$, $\eta_p^2 = .22$, confirmed that while the improvement among the AFRS-High ($p_{bonf} < .001$) and AFRS-Low ($p_{bonf} < .001$) conditions differed significantly from the control condition, they did not differ from each other ($p_{bonf} = .166$). Therefore, the difficulty of the

trial that the AFRS erred on (high vs. low human accuracy) did not modulate the increase in aided performance at test.

Accuracy

Match Trials. The main effect of Task Phase was significant, $F(1, 127) = 65.25$, $p < .001$, $\eta_p^2 = .34$, as was the main effect of Aid Condition, $F(2, 127) = 3.21$, $p = .044$, $\eta_p^2 = .05$, and their interaction, $F(2, 127) = 8.29$, $p < .001$, $\eta_p^2 = .12$ (see Figure S3). Match trial accuracy increased at test for participants in the AFRS-High ($F = 37.13$, $p < .001$) and AFRS-Low conditions ($F = 34.07$, $p < .001$), but not the control condition ($F = 2.50$, $p = .121$).

Mismatch Trials. The main effect of Task Phase was non-significant, $F(1, 127) = 0.55$, $p = .460$, $\eta_p^2 = .00$, as was the main effect of Aid Condition, $F(2, 127) = 1.12$, $p = .329$, $\eta_p^2 = .02$. However, the interaction between the two factors was significant, $F(2, 127) = 5.00$, $p = .008$, $\eta_p^2 = .07$ (see Figure S3). The significant interaction arose from there being no change to the mismatch accuracy of the AFRS-High ($F = 2.22$, $p = .145$) or AFRS-Low conditions at test ($F = 3.57$, $p = .065$), but a significant decrease in accuracy for the control condition ($F = 5.27$, $p = .027$).

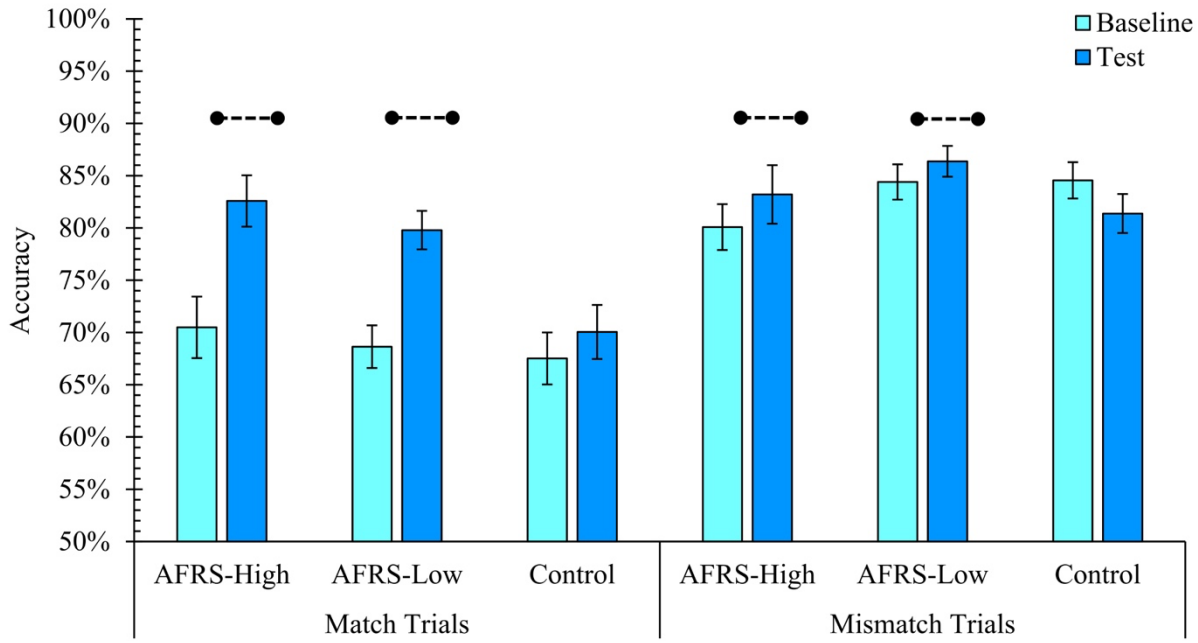

Figure S3. Accuracy for match and mismatch pairs in the baseline and test task phases, plotted separately for each Aid Condition (AFRS-High, AFRS-Low, control) in Experiment 3. The dashed bars show the performance of the AFRS alone. All error bars show the SEM.

Criterion

The conservative response bias was significant for all three aid conditions at baseline (see Table S3). This bias remained significant at test for the AFRS-Low and control conditions, but not the AFRS-High condition.

Table S3

One sample *t*-tests comparing the response bias shown by each Aid Condition (AFRS-High, AFRS-Low, control) in Experiment 3 to 0, in order to determine whether the response bias was statistically significant.

Aid	Task Phase	*M (SD)*	*df*	*t*	95%CI	*p*	*d*
AFRS-H	Baseline	0.17 (0.46)	37	2.35	0.02, 0.33	.024*	0.38
	Test	0.03 (0.52)	37	0.33	-0.14, 0.20	.746	0.05
AFRS-L	Baseline	0.31 (0.38)	46	5.56	0.20, 0.42	< .001*	0.81
	Test	0.14 (0.35)	46	2.63	0.03, 0.24	.012*	0.38
Control	Baseline	0.32 (0.48)	44	4.54	0.18, 0.47	< .001*	0.68
	Test	0.19 (0.50)	44	2.59	0.04, 0.34	.013*	0.39

*Identifies statistically significant comparisons.

Image Pair Analysis

Here we report the effect of the AFRS decision label (correct, incorrect, control) on the accuracy for each face pair, separately for match and mismatch trials. As reported in the main text, accuracy generally increased across all image conditions when the AFRS gave the correct decision (see Figure S4). Conversely, an incorrect decision from the AFRS generally resulted in decreased accuracy across all trial types. The change in accuracy for the "hard" errors appears to be more extreme than for the "easy" errors, which is consistent with the suggestion that participants were less certain about the answer to these difficult pairs and instead relied on the decision-aid. For both correct and incorrect AFRS decision labels, the change in accuracy generally exceeded that of the control condition, which was less than 3% in each label condition, demonstrating that the AFRS decision labels had a significant effect on human responses.

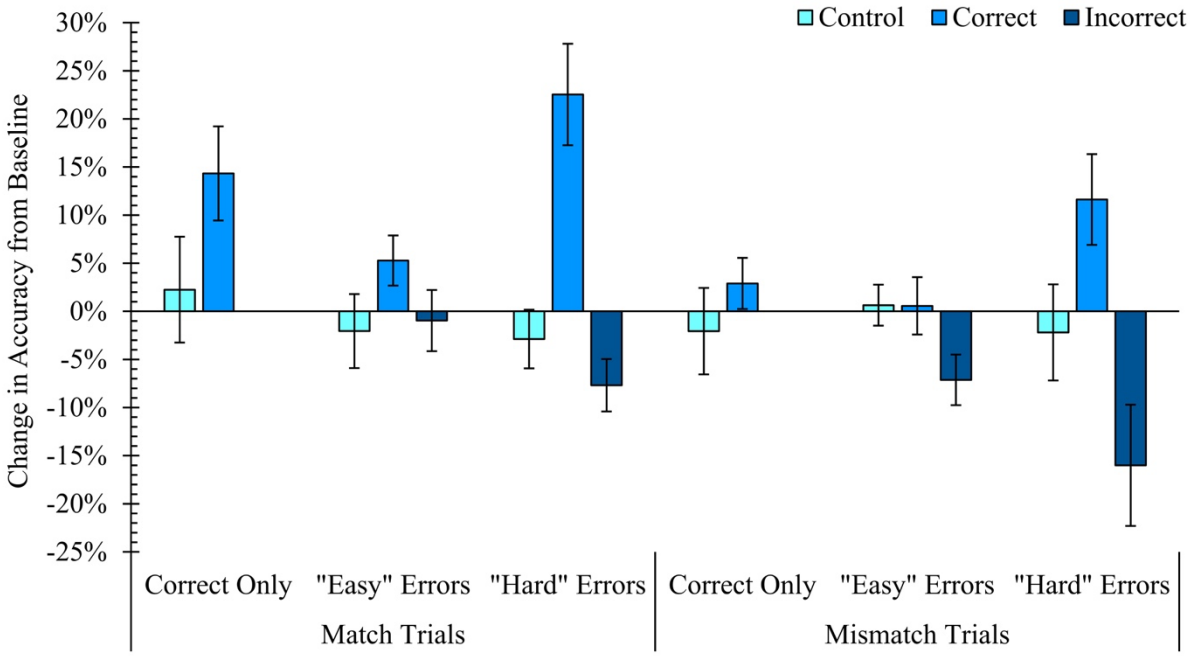

Figure S4. The average change in accuracy from baseline for items that were always shown with the correct label (n = 136), those that were the "easy" errors (n = 16) for AFRS-High, and those that were "hard" errors (n = 16) for AFRS-Low, plotted separately for match and mismatch trials (n/2), when shown with correct, incorrect and control AFRS decision labels. Error bars show 1SD around the mean.