Optimizing Face-matching with Artificial Intelligence (AI) through Trust Calibration

Yu Wa Ng

Submitted for the degree of Doctor of Philosophy (PhD)

School of Psychology & Neuroscience

College of Medical, Veterinary & Life Sciences

University of Glasgow

March 2024

Abstract

Face-matching is an important task that is used as an identity verification method in many applied settings. Advances in Artificial Intelligence (AI) have meant that facial recognition systems are continuously improving in terms of accuracy and are increasingly incorporated into the workplace. The use of e-gate at the border is an example of this technology and illustrates the involvement of humans in the identity verification process. The overarching aim of this research was to identify and better understand factors that influence this human-AI interaction. There was a particular focus on understanding the role of trust, to investigate the possibility of trust calibration. Calibrated trust was expected to help facilitate the interaction and improve face-matching performance. This thesis details the experimental studies that were conducted to achieve this aim.

Chapter 3 presented the findings of Pilot Study 1, providing baseline results of using AI support in a face-matching context by comparing face-matching performance between different groups using AI support of high or low reliability or no AI support. Findings showed that using AI in the decision-making process improved accuracy, particularly when AI was reliable. When AI had limited reliability, this did not affect performance more than not using AI support. The impact of AI errors on human performance was also explored. Experiment 1, also in Chapter 3, used a similar experimental design and found that using AI with limited reliability introduced response bias showing that participants were more likely to believe that face pairs were of matched identities regardless of the truth.

Chapter 4 consisted of Pilot Study 2 and Experiment 2, both designed to examine the influence of presenting AI scores, a quantitative measure of (dis)similarity between two faces, in the face-matching decision process. Pilot Study 2 examined the influence of AI

scores by comparing the performance of participants completing a face-matching task with or without AI scores. Findings showed AI support in the form of (dis)similarity scores influenced performance compared to using no AI. Experiment 2 showed that AI scores do not help calibrate trust when dissimilarity scores were presented alongside incorrect AI labels and there were no effects on face-matching decisions.

In Chapter 5, the final study examined the influence of face-matching expertise on face-matching performance with AI. The experiment recruited face-matching professionals and novices to take part in a face-matching test using AI support. Findings showed both face-matching professionals and novices experienced a decrease in trust and performance on trials where AI provided incorrect advice. The role of confidence in trust was also discussed.

In summary, the results of this research highlighted the impact of using AI on face-matching performance and the role of trust in the use of facial recognition as a decision support system. The thesis concluded with recommendations on the use of AI in a face-matching context and directions for future research are discussed to further the current understanding of human-AI collaboration in face-matching and calibrating trust to facilitate team collaboration.

Table of Contents

Acknowledgements

I would like to thank my supervisor Frank Pollick for his invaluable advice and encouragement throughout this project. Forever grateful for his support and guidance.

I would also like to express my gratitude to Christoph Scheepers for bringing in his insights and expertise and Reuben Moreton for his valuable contributions, feedback and ideas.

A special thanks to Qumodo Ltd., London for their support as well as the Economic and Social Research Council (ESRC) for their funding which made this research possible.

Thank you to members of my lab group who have offered assistance when I needed help.

Last but not least, I would like to thank my family members for their unwavering love and patience and my partner who helped me stay grounded during difficult times along with my pets who have been remarkably attuned to my emotional well-being and offered comfort and companionship at the right times.

Thank you to all who have supported me along the way.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Definitions/Abbreviations

AFR – Automated Facial Recognition

AI – Artificial Intelligence

AUC – Area Under Curve

DCNN – Deep Convolutional Neural Networks

FPR – False Positive Rate

ROC – Receiver Operating Characteristic (Curve)

TPR – True Positive Rate

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background and Context

Matching unfamiliar faces for verification purposes is common in a variety of settings, such as buying age-restricted products, accessing services or premises, forensic investigations and border control. Despite being widely used in many contexts, unfamiliar face matching is susceptible to errors, and error rates could reach as high as 30% (Megreya & Burton, 2008). Previous research in the area of face-matching has identified different factors that reduce accuracy and explored ways to improve face-matching performance, for example, by examining whether different forms of training could be useful in improving face-matching abilities (White, Kemp, Jenkins, & Burton, 2014), or more recently recruiting super recognisers who have extraordinary abilities in remembering and matching faces (Bobak et al., 2016). The current research focused on looking at how face-matching performance could be improved using Artificial Intelligence (AI).

With the development of technology and AI, identity verifications can be automated, as illustrated by the use of e-gates and other similar applications. However, it is also recognised that AI is not fully dependable. For instance, accuracy varies across race and gender (Albiero et al., 2022), and AI is equally susceptible to fraud attempts compared to humans (Robertson et al., 2017). In addition to technical failures, there are many reasons to require the need for humans to be involved in the decision-making process. The task of the human may involve monitoring the system or making a final decision by considering the output of a facial recognition algorithm (Sanchez del Rio et al., 2016). For both practical and ethical considerations, there is a need for efficient and effective human and AI interaction.

Making use of the idea of trust in AI in the human-factors literature, an objective of the current research is to examine ways to optimise and facilitate human-AI interaction. Trust in AI is a concept that is underexplored in the context of face-matching, despite its relevance in determining whether and how technology is used (Parasuraman & Riley, 1997). The current thesis therefore makes this connection by examining the role of trust in face-matching performance and exploring trust calibration as a potential pathway to enhance human-AI interaction.

1.2 Scope and Focus

Due to the interdisciplinary nature of the topic, extensive research was drawn from fields of psychology, human-factors and AI-related literature. To bridge the gap between the psychology of face-matching and trust in AI, this research brought together research on relevant studies involving automation, decision aids, support systems and other types of tasks similar to face-matching. In particular, classification tasks involving perceptual and cognitive processes to make a categorical decision appeared to provide useful insights.

Increasingly, automated systems are becoming more autonomous, which suggests that systems are transitioning to be more independent and capable of learning, producing possibly unexpected outcomes (Hancock, 2017). The thesis examined how AI-based facial recognition can be used in an automated way, but it was recognised that the findings of current research may also apply to autonomous technology. It is also important to distinguish between robotic AI, virtual AI and 'embedded' AI, as the trajectory of trust differs based on the tangibility of AI (Glikson & Woolley, 2020). The focus of the current project was to answer crucial questions on the interaction between humans and AI and more on the task of face-matching and trust, rather than the impact of an AI classifier's degree of autonomy.

It was recognised that the general trust and acceptance of facial recognition technology varies depending on its specific application and across individuals, with examples of concerns such as its potential privacy intrusions, bias, storage and security of data and potential covert use (Kostka & Meckel, 2021). The thesis focused more on trust from the user's point of view, with a specific focus on one-to-one face matching for the purpose of identity verification. From this perspective, the aim was to calibrate trust to reduce bias and improve performance, as opposed to improving acceptance of facial recognition technology. The emphasis is to match a user's level of trust with the AI's actual capabilities via trust calibration, the process of matching a user's level of trust with the actual capabilities of AI (Lee & See, 2004), rather than simply improving or reducing overall trust towards the system.

1.3 Research Aims and Objectives

The overarching aim of the thesis was to explore how human operators making face-matching decisions can be optimally assisted by face-matching algorithms through trust calibration. This research was driven by practical motivations, with the intention of improving face-matching accuracy in applied settings, but also has theoretical significance by verifying the role of trust in this specific context. As the concept of trust in AI in face-matching is underexplored, the current research contributes to the existing literature by examining trust and its influence on face-matching decisions. There were several research questions designed with this aim in mind, examining the impact of using AI in face-matching, the influence of AI reliability and AI scores on face-matching and trust, and the role of expertise in human-AI interaction.

1.4 Key Terms

Face matching involves comparing two faces presented simultaneously and deciding whether they belong to the same person or different people. Past research has focused on the task of face-matching, involving the comparison of faces presented simultaneously. This is in contrast to face recognition, which involves the presentation of stimuli and a subsequent test of memory (e.g. Brown, Deffenbacher & Sturgill, 1977). The thesis also focused on one-to-one face-matching, which involves comparing two faces only. This is in contrast to selecting a face from a list of candidates (Heyer et al., 2018), also referred to as one-to-many face-matching.

Facial recognition technology has shown improvements in terms of accuracy but many systems involve a human in the loop. For example, at the border, individuals who cannot be identified by automated facial recognition systems are transferred to human officers who have to carry out the identity verification manually. The overarching aim of this thesis was to investigate ways to combine the decisions of human operators and facial recognition algorithms to improve face-matching performance by examining the human-AI interaction. Therefore, the focus is less on the AI model used, and more on the interface and interaction.

One way to investigate how to facilitate the interaction is through the examination of trust. Interpersonal trust is an abstract concept that involves cooperation, risk and vulnerability (Rotter, 1967). There are many definitions of trust, adapted in a way relevant to a specific field. It is recognised that there are both similarities and differences between human-human trust and human-AI trust (Madhavan & Wiegmann, 2007), and therefore the

current thesis focused primarily on trust in AI, by examining the relevant influencing factors within the areas of dispositional, learned and situational trust (Hoff & Bashir, 2015).

1.5 Thesis Overview

This research was driven by practical motivations, with the intention of improving face-matching accuracy in applied settings, but also holds theoretical significance by investigating the role of trust in this specific context. The thesis will begin with a review of the literature on face-matching and facial recognition systems, focusing on their interaction and the concept of trust (Chapter 2). Chapter 3 consists of Pilot Study 1, a pilot study designed to clarify the impact of using AI in the face-matching process and Experiment 1 which focused on face-matching performance when presented with AI of high or low reliability. There will be a closer examination of trust in Chapter 4 by looking at the usefulness of presenting dissimilarity scores on face-matching performance in Pilot Study 2 and whether it adds transparency to AI in Experiment 2. Chapter 5 details the final experimental study which looked at the role of expertise in face-matching performance with AI and its implications on trust and confidence. Chapter 6, the final chapter, summarises the key findings and presents recommendations for future research to explore the remaining unanswered questions.

Chapter 2: Literature Review

2.1 Matching unfamiliar faces

Personal familiarity plays a crucial role in moderating performance in face recognition. Prior research on face perception has revolved around the debate on whether humans qualify as face experts. The criteria for expertise have also been a subject for discussion but it is generally agreed that humans possess highly specialized skills and are experts in perceiving and recognizing faces, but only for familiar faces (Young & Burton, 2018). When asked to match high-quality images to faces seen in video clips, students were much more able to recognize lecturers who have taught them previously, even if the face or body of the image was obscured, compared to students who were not familiar with the target to be identified (Burton, Wilson, Cowan & Bruce, 1999). This familiarity-based benefit appears to be a result of repeated exposure to different variations of the same face which contributes to the learning of a new face (Clutterbuck & Johnston, 2005). Exposure to the same face under different illuminations and circumstances averages together to create view-independent familiar face representations that aid recognition despite changes in expression (Johnston & Edmonds, 2009). Our abilities with recognising and matching unfamiliar faces are rather poor compared to familiar faces as they are processed qualitatively differently.

The process of unfamiliar face matching, particularly in contexts such as forensics and border security where errors can have severe consequences, has received significant research attention. Earlier studies initially focused on face recognition tasks, which involve remembering a previously seen face (e.g. Brown, Deffenbacher & Sturgill, 1977). However, it became evident that understanding abilities in face matching which involves matching a person's face with a photograph for identity verification is also an area worthy of research

attention. Research suggests that unfamiliar face matching is equally poor compared to unfamiliar face recognition even with the memory component removed, unless the unfamiliar faces were inverted thereby disrupting the configuration of the face (Megreya & Burton, 2006). When asked to sort real-life photographs of faces into piles based on identity, unfamiliar viewers frequently perceived images of the same individual as different, while familiar viewers performed the task relatively accurately (Jenkins, White, Montfort & Burton, 2011). This demonstrates that both between-person and within-person variability contribute to the difficulty in matching unfamiliar faces.

The highly error-prone task of unfamiliar face matching is particularly problematic in applied contexts such as in forensic settings and at the border where errors could lead to severe consequences and a matter of national security. Various face-matching tests have been developed to assess unfamiliar face-matching abilities. The Glasgow Face Matching Test (GMFT) is a test designed specifically for unfamiliar face matching and involves matching pairs of faces taken in the same full-face view (Burton, White & McNeil, 2010). Images from the GMFT were taken with different cameras which introduced a degree of variability, increasing the difficulty of the task. The GMFT has been used in many subsequent studies of unfamiliar face matching. Another popular option appears to be the Kent Face Matching Test (KFMT) which had been designed to be more ecologically valid in applied settings, using well-lit photographs and student ID photos that were taken months apart under unconstrained conditions with varying expression and pose across individuals (Fysh & Bindemann, 2018). These tests have been used in many studies as a validated assessment of unfamiliar face matching and highlight the difficulty of the task. More recently, an expansion of the GFMT has been developed. The GFMT2 included more image and person variations which made the test more difficult and representative of

everyday tasks (White et al., 2021). The GFMT2 also contains different versions of the test designed for exceptionally high-performing or low-performing individuals to support research on individual differences in face-matching.

2.1.1 Performance-related Factors

It is evident that various aspects, including individual differences, image-based factors, and observer-related elements, can significantly impact performance in face-matching, which highlights the need to consider and control these variables in face-matching studies. Using benchmark tests of unfamiliar face-matching such as the GFMT and KFMT, many researchers have found a number of factors that could impair face-matching performance. The different ways that individuals vary could result in reduced speed and accuracy in face matching. For example, images of faces wearing glasses can create enough variability to impair face-matching performance as participants become more conservative and cautious with their decisions (Kramer & Ritchie, 2016). Image-based factors, relating to the way images are taken and the general quality of the stimuli can also be a source of variability. For instance, reduced image quality by removing pixels affects unfamiliar face matching, substantially more than unfamiliar face recognition (Bindemann, Attard, Leach & Johnston, 2013). Even subtle alterations such as putting pictures in a passport frame also reduced participants' abilities to detect a mismatch (McCaffery & Burton, 2016). Factors relating to the observer could impair performance further, such as differences in ethnicity between the viewer and the subject in the face image (Megreya, White & Burton, 2011), and the gender of the viewer (Herlitz & Lovén, 2013). Performance varies between individuals but also within the same viewers as some observers make different decisions on the same faces on different days (Bindemann, Avetisan & Rakow, 2012). These experiments

illustrate the factors that must be considered and ideally controlled for in face-matching studies.

In real-world scenarios such as border control, error rates are expected to be higher. This can be attributed to the unequal ratio of identity matches and mismatches, with identity mismatches often undetected in low-prevalence conditions compared to high-prevalence conditions (Papesh & Goldinger, 2014). The combined effects of time pressure and the repetitive nature of the task are also factors that could further reduce face-matching accuracy (Bindemann et al., 2016). Regular five-minute rest breaks with entertainment provided and desk switching do not appear to eliminate this decline (Alenezi, Bindemann, Fysh & Johnston, 2015). Research therefore needs to be applicable to face-matching scenarios outside of the laboratory as errors in these applied settings carry higher stakes and consequences.

Understanding the role of experience in face-matching has also been a focal point of investigation. Many studies have been carried out to explore face-matching abilities in a variety of populations but research has shown mixed findings. For instance, a large-scale online study that compared unfamiliar face-matching performance of notaries, bank tellers and undergraduate students found no correlation between experience and performance (Papesh & Goldinger, 2018). Student performance has also been compared to groups with more relevant backgrounds such as police officers and passport officers. For example, forensic examiners with many years of experience comparing face images have been shown to have their performance improved compared to untrained students but only at long exposure durations, possibly indicating the usage of previous formal training (White, Phillips, Hahn, Hill & O'Toole, 2015). Similarly, passport officers were shown to have a slight advantage at matching photos to official IDs compared to students but also took

significantly longer and further analyses indicated that the length of employment did not predict accuracy (White, Kemp, Jenkins, Matheson & Burton, 2014). Police officers appeared to be better than novices at matching unfamiliar faces, but their experience was also not related to accuracy (Wirth & Carbon, 2017). These studies highlight the complex relationship between experiences and face-matching accuracy.

A more robust phenomenon found in the literature is that face-matching tasks seem to be approached differently by experts and novices. For example, one study found that untrained students had more false positives and false negatives than forensic experts, as experts tended to be more careful with their conclusions when image quality was low (Norell, Lathem, Bergstrom, Rice, Natu & O'Toole, 2015). Experts required detail in the image to perform the face-matching task accurately. Furthermore, untrained participants used more non-face identity information than facial examiners who also used the rating scales differently resulting in fewer false positives (Hu et al., 2017). These findings confirm the differences in approach between novices and experts but how these differences could translate to higher face-matching accuracy would require further research.

Other studies have clarified that experience, as opposed to training, in facial analysis may be an indicator of performance in the ability to identify faces from CCTV footage as facial image analysts with at least five years of professional experience were significantly better than non-experts (Wilkinson & Evans, 2009). However, it remains unclear whether these experts had an innate ability in face matching, and the extent of the training was not specified. This aligns with the broader finding that there are large variations in individual performance in face matching and memory for faces in general, both of which appear to share the same mechanisms (Fysh, 2018). The distinction between novices and experts in

face-matching strategies is apparent, but further research is needed to explore ways to enhance face-matching accuracy.

Exploring the impact of expertise also involves examining the face-matching abilities of superrecognizers and the potential benefits of training. Superrecognisers are individuals with extraordinary face recognition and perception abilities who perform significantly better than average people on standardised face memory tests (Russell, Duchaine & Nakayama, 2009). Superrecognisers have been found to perform consistently better than students on both unfamiliar and familiar face-matching tests (Robertson, Noyes, Dowsett, Jenkins & Burton, 2016). Similar to forensic experts, these people are also more conservative with their decisions as they are more likely to reject images as mismatches. Their abilities are independent of motivational levels as superrecognizers perform better than controls regardless of whether or not they have monetary incentives (Bobak, Dowsett & Bate, 2016).

Research on the effects of training is less conclusive. Face shape training by asking participants to classify unfamiliar faces according to their face shape (e.g. oval, round, pear etc.) has failed to improve matching accuracy in the GFMT (Towler, White & Kemp, 2014). On the other hand, an extensive training program with participants learning 30 different identities using multiple images of each identity over several days enhanced performance for matching new images but this effect did not generalise to untrained identities and training using a single image of each identity did not have this effect (Matthews & Mondloch, 2018). Providing instructions for student participants to attend to specific facial features enhanced accuracy but only for certain features (Megreya & Bindemann, 2018). Facial image comparison training which involved studying the facial anatomy and morphological comparison yielded no improvement in identification accuracy, despite participants believing that they had improved, but longer (3-day) courses using a feature-by-

feature comparison strategy produced better results (Towler et al., 2019). Differences in on-the-job training and opportunities for deliberate practice may be a feasible explanation for these mixed results. Further research could tease apart the different components of a training program that may be more effective than others, assuming that prior training contributes to the idea of expertise.

The role of feedback on facial recognition has also been examined. A series of experiments have repeatedly found a beneficial effect of feedback in maintaining but not improving accuracy in mismatch trials. Specifically, the type of feedback given appears to be of importance, with trial-by-trial feedback producing this beneficial effect but not when feedback was given at the end of the block (Alenezi & Bindemann, 2013). Another study has shown that training by providing trial-to-trial feedback benefits simultaneous face-matching with the effect of generalising to novel, unfamiliar face images (White, Kemp, Jenkins, & Burton, 2014). This study compared a feedback group with a no-feedback group, effectively separating the effects of feedback and mere practice. Feedback possibly helps improve accuracy by increasing self-awareness as participants tend to have a poor understanding of their face-matching abilities for unfamiliar faces (Bindemann, Attard & Johnston, 2014). Working in pairs where discussions were allowed also improved overall accuracy, in addition to subsequent individual performance (Dowsett & Burton, 2015), providing a route to training.

2.2 Facial Recognition Technology

The integration of AI into face-matching has both practical and theoretical considerations. In response to the challenges of face-matching, one approach is to replace human face-matching decision-makers entirely with AI systems. However, an alternative and more promising strategy involves the teaming of humans and AI, providing even more

accurate outcomes. Facial recognition is used for identification or verification purposes with many applications in information security, surveillance, access control and law enforcement. Its increasing popularity and research attention are perhaps driven by its applications in identification and verification systems. Identification involves matching an image to one that is already stored in a database, also known as one-to-many face-matching. On the other hand, verification involves matching two images based on their degree of similarity thereby confirming that the two faces belong to the same person, also referred to as one-to-one face-matching. An application of verification matching is at the borders using e-gates where a person presents an official document like their passport to an automated facial recognition system (AFR) which compares the image to the live person. This context highlights the practical significance of implementing facial recognition technology.

Even when dealing with high-quality images, the accuracy of algorithms is limited by factors such as lighting, pose, expression and location (Beveridge et al., 2011). The development of Deep Convolutional Neural Networks (DCNNs) models, trained with a large number of diverse images spanning multiple variables including age, makeup, hairstyle, facial hair and glasses, addressed the challenges of unconstrained images (O'Toole, Castillo, Parde, Hill & Chellappa, 2018). DCNNs are able to recognise faces across viewpoints, illumination, expression and appearance as they create a unitary space that captures both facial identities and face images (O'Toole et al., 2018). Independent evaluations by the US government such as the FERET, Face Recognition Vendor Test (FRVT) and Multiple Biometrics Evaluation (MBE), administered by the National Institute of Standards and Technology (NIST) provide a good benchmark of performance over the past years. There was a rapid decline in error rates between the years 1997 to 2010 (Phillips, 2011). Current AFR systems work optimally under controlled conditions but do not perform comparatively

to humans in unconstrained environments with variations in pose, illumination, ageing, occlusion, expression, plastic surgery and low resolution (Oloyede, Hancke & Myburgh, 2020).

Research directly comparing human and algorithm performance has revealed that out of 7 state-of-the-art face recognition algorithms used in the Face Recognition Grand Challenge (FRGC), held by the U.S. Government to advance face recognition technology, six surpassed humans on face pairs pre-screened to be 'easy' and three surpassed humans on face pairs pre-screened to be 'difficult (O'Toole et al., 2007). Another study found that humans performed better with unlimited viewing times than with short exposures to pairs of faces but performance in both conditions was inferior to algorithm performance using the top-performing algorithms from the FRVT 2006 for 'good' and 'moderate' images and were only comparable to algorithm performance for 'poor' images (O'Toole, An, Dunlop, Natu & Phillips, 2012). These images were taken from the biometric data set from the FRVT 2006 challenge and explicitly controlled for influential factors such as subject age, pose, and changes in camera by capturing images in the same academic year, using the same model of camera to take full frontal face images and using an image of the same person in each of the categories. Images divided into 'good', 'bad and 'ugly', corresponded to easy, medium-level and difficult-to-match face pairs (Phillips et al., 2012).

Algorithms from the FRGC have shown different sensitivities to images of subjects wearing glasses and performance varies for different expressions (Beveridge et al., 2009). There have been suggestions that some algorithms may be biased against certain races that were not used in the training data (Furl, Philiips & O'Toole, 2002) and are affected by other demographic covariates as accuracy tends to be greater for males than females and older people than younger people (Abdurrahim, Samad & Huddin, 2018). Facial recognition

models and current off-the-shelf commercial algorithms are vulnerable to imposter faces that are underexposed or overexposed with adjusted levels of brightness, leading to higher false match rates, which could potentially explain differences in accuracy across demographic groups (Wu et al., 2023). Even if accuracy is high, algorithms may still be vulnerable to mask-spoofing attacks (Kose & Dugelay, 2014). Together these favour a human-in-the-loop approach where a human operator is required to verify or rectify a decision made by an AFR system to maintain higher levels of accuracy and security.

2.3 Human-AI Interaction

The wisdom-of-crowds effect (Surowiecki, 2004) for face recognition in human participants has been previously demonstrated, where the grouping of response data made independently by untrained non-experts enhanced accuracy in the GMFT compared to working alone (White, Burton, Kemp & Jenkins, 2013). Applying this idea, research also found that fusing algorithm scores with human judgements substantially improved performance in a 'difficult' face-matching task (O'Toole, Abdi, Jiang & Phillips, 2007). This approach attempts to take advantage of the different strategies used by humans and AI. For example, untrained students are able to use information from the body when the face does not contain enough identity information while facial examiners are more likely to focus only on the face (Hu et al., 2017). With experience and individual differences taken into account, a more recent study comparing the performance of professionals and superrecognisers with undergraduate students and algorithms has revealed that fusing the highest-performing group with the best-performing algorithms yielded the greatest level of accuracy (Phillips et al., 2018). Collaborative efforts achieve higher accuracy than humans or algorithms alone.

However, this partnership between humans and machines could introduce further issues and bias. More recently, there has been research focusing on the effect of this

interaction on face-matching accuracy. In an identification task using real-life passport photographs, facial reviewers asked to compare an image to a candidate list of possible matches made an error on average in 1 of every 2 candidate lists and were no better than untrained students (White, Dunn, Schmid & Kemp, 2015). After being provided with a target image, an AFR system returns a candidate list of images, ordered and ranked by their degree of similarity. The number of candidate matches presented to reviewers significantly affected performance in an unfamiliar face-matching task. The study found that longer lists with 100 images produced more false alarms, lower confidence ratings and increased response latencies in both experienced and inexperienced facial reviewers (Heyer, Semmler & Hendrickson, 2018). This is similar to the findings in research on fingerprint examiners and automated fingerprint identification systems (Dror & Mnookin, 2010). Difficulties are expected to increase as algorithms select more challenging matches that are similar to each other and the target. The user interface for automated systems therefore requires careful design to reduce bias and to aid the decision-making process in a meaningful way.

Other studies have shown that prior decisions made by algorithms influenced the subsequent face-matching decisions made by human operators. When face pairs were inconsistently labelled as ‘same’ on a mismatch trial or ‘different’ on a match trial, accuracy decreased by drawing attention away from the face images (Fysh & Bindemann, 2018). Mimicking the higher frequency of matches to mismatched cases at the border, this study illustrated the potential for errors in human-computer interaction at passport control and suggests that text cues may be detrimental to performance when they are inaccurate. Likewise, another study has shown a similar bias as participants were mostly correct when no prior information was given but introducing labels biased their certainty judgements (Howard, Rabbitt & Sirotin, 2020). The study made use of signal detection

theory (Stanislaw & Todorov, 1999) to examine participants' ability to discriminate between faces and confirmed that participants were biased cognitively, without changes in sensitivity to the similarity of faces. Another study, also focusing on face-matching and interactions with AI, has found that participants had higher sensitivity and trust in AI in scenarios where the human makes the first decision, and lower sensitivity and higher trust in cases where the AI makes the first decision (Salehi et al., 2021). These studies show that using AI in face-matching has a direct influence on performance.

Effects of the interaction between humans and AI in one-to-one face-matching appears to have been examined to a very limited extent but similar findings have been reported in other fields. Research in healthcare AI and clinicians has also examined human-AI collaboration, specifically exploring whether this collaboration is affected by the way in which information from the AI is presented. It was found that using faulty AI where output was manipulated in a way that favoured an incorrect diagnosis, the performance of clinicians deteriorated, even changing their initial decision in diagnosing skin cancer (Tschandl et al., 2020). A systematic review of research on diagnostic performance and machine-learning-based decision support systems found no concrete evidence to suggest that using AI improves decision-making in clinical settings (Vasey et al., 2021). This highlights that the collaboration between humans and AI introduces risks and problems that may not be specific to face recognition.

2.4 Trust in AI

Trust is an important factor to consider in human-AI interactions, as the occurrence of human errors can be attributed to a mismatch of trust between humans and AI. Miscalibrated trust can lead to inappropriate use of automated technology, with inappropriate use conceptualised as instances of misuse, disuse or abuse of technology

(Parasuraman & Riley, 1997b). Trust calibration therefore seems to be an appropriate way to facilitate human-AI interaction between face-matching decision-makers and facial recognition algorithms.

2.4.1 Importance of Trust

Trust has been examined in a wide range of fields, with many different conceptualisations. Earlier research has framed trust as an attitude or expectation of a favourable response in a cooperative relationship (Rotter, 1967). In contrast, trust has also been conceptualised as a belief in the trustee's ability, benevolence and integrity, which involves an intention, the willingness to take risks and an element of vulnerability (Mayer et al., 1995). Furthermore, trust can also manifest a behavioural outcome such as compliance (Meyer et al., 2014). Studies have employed different definitions and measures of trust depending on whether the trust was conceptualised as a psychological state or choice of behaviour (Kramer, 1999). From a social perspective, trust is a cognitive process involving the estimation of risk and a decision on whether to rely or not on an agent (Castelfranchi & Falcone, 2000). These illustrate the wide range of approaches to trust across different disciplines.

To address potential inconsistencies between different definitions, Lee and See (2004), in an influential integrated review of early research on trust and reliance on automation, utilised Fishbein and Ajzen's (1975) Theory of Reasoned Action and proposed that trust is an attitude that leads to the intention of a behaviour. The defining and understanding of trust remain the subject of extensive ongoing debate (Costa, Fulmer, & Anderson, 2018; Eikeland & Saevi, 2017), but Lee and See's (2004) examination of trust and its effect on reliance on automation has had great influence in the research fields of trust in automation.

Early research has noted that humans interact with computers in a similar way to how they would interact with other humans in a cooperative relationship. A study using laboratory-based games found that humans viewed computers as teammates and were more open to influence from the computer (Nass et al., 1996). In addition to that, automation that has person-like characteristics can influence trust and increase dependence on the automated aid (Pak et al., 2012). In line with this, research has found human-human trust to be comparable to human-automation trust, with important similarities such as positivity bias in trusting novel technologies (Madhavan & Wiegmann, 2007). This highlights that trust is a relevant and valid concept in contexts involving human-AI interactions.

Trust has long been argued to be the key to mediating the human-automation relationship (Muir, 1994). Trust along with other constructs such as mental workload and situational awareness has been recognised as a predictor of human-system performance (Parasuraman et al., 2008). Trust combined with confidence influences interactions with automation. For example, when trust exceeds self-confidence, automation is used but when confidence exceeds trust, manual control is used (Lee & Moray, 1994).

Miscalibrated trust appears to have negative consequences. Depending on the role of the human operator, consequences could be automation-induced complacency or automation bias (Parasuraman & Manzey, 2010). One important aspect of automation misuse is reflected in insufficient monitoring or checking of automated functions, a phenomenon which commonly has been referred to as complacency (Parasuraman, Molloy & Singh, 1993), which is often discussed in the context of supervisory control. Complacency appears to be most relevant in monitoring tasks involving attention (Moray & Inagaki, 2000). The errors that arise could be a result of attentional bias or discounting contradictory

information (Manzey et al., 2012). The impact of automation on human performance has been widely studied and includes many challenges such as the loss of situational awareness, and effectively explaining why human operators do not take control when needed (Endsley, 2017).

Of particular relevance is automation bias, which is related to the way humans use the outcomes of an automated decision, given by support decision aids. This leads to one of two behavioural outcomes, which are omission and commission errors (Bahner, Elepfandt, & Manzey, 2008). Omission error is when the human operator's over-reliance on the automation results in a failure to notice an automation error if the automation does not alert them to it and commission error is when the human operator follows recommendations given by the automation, despite the recommendation being wrong (Skitka et al., 1999). A higher cognitive load in more complex tasks is more likely to increase automation bias errors (Lyell et al., 2018). Calibrating trust between human operators and facial recognition systems may initially involve examining the types of errors made by humans. Automation-induced errors are not specific to face-matching and facial recognition and are evident in many applied settings, including healthcare and medicine (Jacobs et al., 2021; Goddard et al., 2012). For example, the use of decision aids in tasks within these settings is conceptually similar to face-matching, requiring both perceptual and cognitive processes along with categorical decisions.

2.4.2 Factors Related to Trust

One model conceptualising the variability of trust in automation includes the human operator, environment and system, reflecting dispositional trust, situational trust and learned trust (Hoff & Bashir, 2014). Dispositional trust is related to individual differences,

such as age, gender and personality while situational trust is influenced by context-dependent factors and learned trust is a product of previous evaluations of a system from experience. These are factors that could influence trust towards automation, which can also be categorised as the environment, the operator and the machine, with the performance of the system having the greatest association with trust (Hancock et al., 2011). Several other reviews have similar classifications, namely person-related variables such as personality, and expertise, system-related such as reliability and situation-related such as workload and affect (Schaefer et al., 2016). These frameworks serve as a useful guide to understanding trust and designing systems that encourage appropriate trust.

Trust varies based on individual experience with a system. Witnessing errors appears to be particularly detrimental to trust levels, and subsequently performance. For example, reduced system reliability results in lower levels of trust toward automation, with reduced speed to compensate for similar levels of accuracy (Chavaillaz et al., 2016). Participants who had initially rated an automated decision aid to be trustworthy lost trust in the aid after seeing errors made by the aid (Dzindolet, Peterson, Pomranky, Pierce & Beck, 2003), indicating that trust is susceptible to individual experiences with automation and can fluctuate accordingly. Trust toward automation before seeing errors tends to be positive. Changes in trust are likely a process of learning. Learned trust is dynamic, changing over time and updated by observations and experience with a system (Kraus et al., 2020). A visual detection task has demonstrated that participants can have a perfect automation schema, as participants tended to have a more favourable bias toward automation than humans but were less likely to rely on an automated system because they noticed and remembered the errors made by AI (Dzindolet et al., 2002). Task difficulty can influence rates of agreement with automation as individuals are more likely to use automation in

trials that were perceived as difficult, and when the task was framed to be important (Schwark et al., 2010).

There are also consequences of losing trust. Under decreasing system reliability, operator performance deteriorates in a number of different contexts such as flight simulation (Bailey & Scerbo, 2007) and self-driving car navigation (Ma & Kaber, 2007). Failures that occur early in an interaction may have very damaging effects (Manzey, Reichenbach & Onnasch, 2012). Trust is harder to repair than to build (Sauer, Chavaillaz & Wastell, 2015) and only recovers slowly over time after exposure to error-free performance (Lee & Moray, 1994). A review of the trust repair literature has suggested a theoretical framework supporting the interaction between trust repair strategy and failure type in human-automation interactions (Marinaccio, Kohn, Parasuraman, & de Visser, 2015).

Characteristics of the user tend to refer to stable traits and can play a role in human-AI interaction. Implicit attitudes toward automation influence trust (Merritt et al., 2013). Propensity to trust can be referred to as an individual's general tendency to trust automation, regardless of the context or the specific system used (Hoff & Bashir, 2015). This predisposition to trust is distinct from the intention to trust (Gill et al., 2005), however, both of these can influence behaviour which makes it difficult to tease apart. Propensity to trust can be measured using self-report methods. Context-specific measures have tended to better predict the perceived trustworthiness of the system and trust behaviour (Jessup et al., 2019).

Propensity to trust interacts with other variables of trust to influence trust ratings. For example, individuals with a high propensity to trust experience a greater decline in trust when given an unreliable AI (Merritt & Ilgen, 2008). A higher propensity to trust in addition

to implicit preferences for automation predicted trust when individuals observe a system making obvious errors (Merritt et al., 2013). This is further supported by findings that show when individuals expect automation to be trustworthy, they tend to be more sensitive to changes in automation reliability (Pop et al., 2015). Expertise plays a role, as novices tend to benefit from the use of automated decision aids more than experts (Chavaillaz et al., 2019). Novice pilots when using automated decision aids display greater complacency potential compared to experts (Lyons et al., 2017).

2.4.3 Trust Calibration

Trust calibration plays a crucial role in ensuring accurate outcomes. Trust calibration could involve perceptual accuracy, perceptual sensitivity and perceived reliability (Merritt et al., 2015). Despite the extensive use of automation and algorithm support in unfamiliar face matching, there is limited research connecting it to the human factor's literature. Specifically, the impact of trust in the face-matching context has been largely unexplored.

Trust calibration is the process of aligning trust with the actual reliabilities of agents (Lee & See, 2004). Research on human operators making face-matching decisions and machines can also examine the process of trust calibration. Research has shown that the humanness of automation can impact trust calibration and influence compliance with an automated aid (De Visser et al., 2012). Explainable artificial intelligence (XAI) has emerged to be useful in aiding trust calibration, as AI predictions accompanied by confidence scores appear to aid trust calibration (Zhang et al., 2020). Research has found that the presentation format of information could make a difference as viewing times tend to be longer for tables than bar graphs. These are information on health records and provide an important way to communicate medical test results to patients (Brewer et al., 2012), emphasizing the

importance of effective communication. In addition to lengthening viewing times, bar graphs can introduce bias as viewers tend to see means lower than they should (Godau et al., 2016). For facial recognition, AI explainability has been shown to be useful in comparing two face images by highlighting areas of similarity between two faces (Lin et al., 2021).

Previous studies have focused on improving system transparency (Yang et al., 2017), often by providing confidence information (McGuirl & Sarter, 2006). Using a behavioural measure of trust to explore ways that improper trust calibration could be mitigated, a study has found that detecting calibration status and presenting cognitive cues to promote calibration during periods of overtrust were much more effective than continuously presenting information on reliability (Okamura & Yamada, 2020). There is a need to examine the user interface and the presentation of information to users of facial recognition systems in relation to trust calibration. Further examining the interaction and improving the interface, will provide insights on how users can be optimally assisted by technology so that quicker and more accurate decisions can be made. To minimise the risk of errors and identification in face-matching tasks, it is important to establish a baseline to provide a minimum level of performance on which to build improvements. The overarching question that the thesis aims to address is therefore how human-AI interactions can be facilitated through trust calibration to reduce errors and enhance performance in face-matching tasks.

Chapter 3: AI and AI Reliability

3.1 Pilot Study 1

3.1.1 Introduction

Face-matching can be a difficult task, raising concerns over this method of verification of identities. Even under optimal conditions, matching faces from images taken on the same day but with different cameras can introduce uncertainty and increase error rates (Burton et al., 2010). In addition to the moderating factor of familiarity, many other variables could affect face-matching accuracy for both naïve participants and trained experts (White, Norell, Phillips & O'Toole, 2017). In applied settings such as security borders, errors could lead to severe consequences. Facial recognition technology can be used to authenticate the identity of a person but often requires a human in the loop to verify the decisions. The current experiment was designed to examine the effect of using Artificial intelligence (AI) in decision-making and explore the impact of AI errors on human performance.

Face-matching is error-prone, which is primarily related to the fact that viewers are often unfamiliar with the faces that they are matching. Research has consistently shown that familiarity with the target is a factor that moderates performance, as unfamiliar viewers perceive images of the same person as different individuals while familiar viewers are more sensitive to within-person variability (Jenkins, White, Montfort & Burton, 2011). Familiar faces, such as famous or previously seen faces are matched more efficiently, including faces that were only seen on brief occasions (Clutterbuck & Johnston, 2010). Familiar faces also tend to be matched accurately unless they are inverted (Megreya &

Burton, 2006). This presents a challenge for individuals whose daily responsibilities might include matching faces that they are not familiar with.

In applied contexts, such as matching faces to passports, additional factors influence accuracy. For instance, time pressure has a detrimental effect on performance, particularly on mismatched trials, and impairs sensitivity (Wirth & Carbon, 2017). In addition to time pressure, time passage is associated with reduced accuracy, as observers tend to show a match response bias over time, where they view two faces as the same identity (Fysh & Bindemann, 2017). These highlight the susceptibility of face-matching to error in real-life scenarios.

Many researchers in the past have compared the accuracy of human and machine performance on face verification tasks and found that algorithms competed quite well, except in challenging tasks where humans performed consistently better (O'Toole, An, Dunlop & Natu, 2012). However, human reviewers have been shown to use non-face information such as from the body and clothes when unable to make a decision solely from the face (Rice, Phillips, Natu, An & O'Toole, 2013). In general, algorithms are superior to humans when matching images that are captured under ideal conditions and illumination (Phillips & O'Toole, 2014). However, algorithm accuracy can be affected by various factors. For instance, real-world scenarios have unconstrained conditions such as illumination and pose variations, occlusion and expressions (Hassaballah & Aly, 2015). Algorithm performance also appears to differ between races as East Asian faces tend to require higher thresholds to achieve the same false-acceptance rates as Caucasian faces (Cavazos et al., 2019). Other factors could include non-demographic attributes that strongly affect recognition performance, such as accessories, hairstyles and colours, face shapes, or facial

anomalies (Terhorst et al., 2021). These findings illustrate that algorithms are not yet perfect and require humans in the loop in the verification process.

Human-AI teams can improve performance. Research has confirmed that fusing the scores made by humans and machines increases accuracy (Phillips et al., 2018), supporting the need for human operators in the decision-making process. This involves considering the output presented by algorithms and making a decision by reviewing two or more images. In the interaction between users of automated facial recognition technology and algorithms, team performance is often limited by human accuracy. For example, accuracy decreases as the number of faces available for comparison increases, resulting in more false alarms and fewer hits in one-to-many identification tasks (Heyer et al., 2018). The way information is presented appears to have a big impact on performance by drawing attention away from the face as inconsistent labels of 'same', 'different' and 'unresolved' reduces accuracy in one-to-one face-matching tasks (Fysh & Bindemann, 2018). Findings examined in light of signal detection theory suggest that the presence of a prior label introduces cognitive bias by impacting the internal threshold that participants have to make a decision (Howard et al., 2020). While these studies indicate that using AI has an influence on face-matching performance, whether the impact on performance is worse than not using AI support at all is less clear. The current experiment included a control condition where no AI support was provided at all to directly ascertain the influence of AI on face-matching performance.

There are potential drawbacks of AI assistance in face-matching when AI provides incorrect recommendations. This pilot study was designed to verify whether the type of error had an impact on performance. Trust tends to be higher when automated aids are prone to false alarms than when the system is prone to misses but systems that are prone

to misses tend to reduce reliance (Davenport & Bustamante, 2010). The type of automation failures experienced during training can also affect performance, as the experience of misses appears to have a bigger impact on error rates than misdiagnoses (Sauer et al., 2016). Another study using a target detection task found that a system prone to false alarms in easy tasks decreased compliance compared to misses as it led to lower trust and more disagreements on difficult trials (Madhavan et al., 2006). The current experiment examined the influence of AI with limited reliability by looking at false positives where matched identities were labelled as different and false negatives where non-matched identities were labelled as the same.

The receiver-operating characteristic (ROC) analysis used in signal detection theory offers a useful insight into the trade-off between hit rates and false alarm rates of classifiers and the area under the curve (AUC) is an indicator of overall performance, with a large total area under the ROC curve indicating a better performance of the classifier (Pintea & Moldovan, 2009). It was hypothesised that by directing the human user towards a correct decision via consistent labels, a reliable facial recognition AI will improve face matching performance compared with using no AI support and that the AUC will be higher in the High AI Reliability group compared to the No AI Support group. The study also intended to explore the idea of AI reliability by distinguishing the impact of the different errors that AI can make. This will be achieved by looking at the effect of using AI that makes errors of labelling matched identities to be different (false positives) and non-matched identities to be the same (false negatives). Thus, when directing users toward an incorrect decision by presenting face-pairs with inconsistent labels, AI with limited reliability will reduce face-matching performance compared to using no AI support and AUC will be lower in the two

Low AI Reliability groups compared to the No AI Support group. However, there will be no difference between the two Low AI reliability groups in terms of AUC.

3.1.2 Methods

Participants.

There was a total of 64 participants recruited through *Prolific* (*www.prolific.com*). Participants were adults aged 18-35, self-identified to be British. Participants were compensated using *Prolific's* payment system, at a rate of £7.00 per hour.

Materials.

Kent Face Matching Test (KFMT): The short version of the KFMT consisted of 40 pairs of faces, each containing one student ID style of image and a high-quality portrait taken at least three months apart (Fysh & Bindemann, 2018). Student ID photos were not controlled by expression, pose or image-capture device.

Face-recognition: This is a Python package available online under MIT license which uses *dlib's* facial recognition algorithms (version 1.3.0.). All KFMT face pairs used for testing were processed by face-recognition and the dissimilarity score for each face pair was obtained, where the lower the score, the higher the similarity between two faces. This model achieved an AUC of 0.99 on the KMFT images and these scores were used to inform us of the labels given to each face pair, in the form of 'same' or 'different' with 12 images with the highest dissimilarity scores labelled as same in the False Positives Group and 12 images with the lowest dissimilarity scores labelled as different in the False Negatives Group.

PsychoPy interface: The experiment is designed on PsychoPy (Peirce et al., 2019). The layout of a single trial consists of a face pair image placed to the left of the screen with a confidence rating slider to the right.

Procedures.

Participants were asked to read the Participant Information Sheet before proceeding to the online experiment hosted on *Pavlovia* (*https://pavlovia.org/*) and to respond to a series of statements regarding consent before beginning. Participants were asked to decide whether two images were the same or different using the confidence statements provided, on a discrete, 7-point rating scale from *"I am absolutely certain this is the same person*" to "*I am absolutely certain these are different people*".

Design.

The experiment used a between-participant design. There was a total of four experimental groups: No AI support, Reliable AI, False Negatives and False Positives. In the No AI Support group, participants were shown face pairs with no AI support. In all other AI groups, a label of 'same' or 'different' was available beneath the face images on each trial. In the Reliable AI group, these labels were fully accurate, reflecting the predictions made by *face-recognition*. In the False Positives group, different identities on non-match trials were labelled as the same and in the False Negatives group same identities on match trials were labelled as different. From the 40 trials, 12 contained inconsistent labels whereby 6 match trials depicting the same identity were labelled 'different' or different identities labelled as 'same'. Randomly allocated in the trials was a face pair depicting a famous politician, included as an engagement check to ensure that participants were following instructions correctly and focused on the experiment.

Method of analysis.

A receiver operating characteristic curve (ROC) is a graphical representation of the performance of a classifier, by taking into account the true positive rate (TPR) and false positive rate (FPR) over a range of possible threshold values, and has applications in psychological research examining an individual's capacity to discriminate between different stimuli (Swets, 1973). The current study used the R package *pROC* to calculate AUC, which represents the area under the ROC curve, a measure widely used to assess and compare the performance of classifiers. Accuracy can be considered low with an AUC between 0.50 – 0.70, moderate with an AUC between 0.70 – 0.90 and high accuracy with an AUC of over 0.90 (Streiner & Cairney, 2013).

3.1.3 Results

There was a total of 64 participants: 25 in No AI Support, 11 in Reliable AI support, 11 in False Negative and 17 in the False Positive group. AUC derived from confidence ratings can provide insights into categorical decision-making (Weidemann & Kahana, 2016). Table 1 is a summary of the mean AUC in each group.

Table 1

Mean AUC for each group

Group	AUC
	M (SD)
No AI Support	0.71 (0.10)
Reliable AI	0.83 (0.14)
False Negatives	0.77 (0.13)
False Positives	0.79 (0.08)

The results of an ANOVA examining the effect of the group on mean AUC were significant ($F(3, 60) = 3.903$, $p = .013$). A post hoc Tukey test showed that the mean difference between No AI Support and Reliable AI was significant ($p = .01$). There were no other significant pairs. Figure 1 displays the distribution of AUC for each group, confirming that Reliable AI has the highest AUC, indicating its ability to distinguish between a match and mismatch trial compared to the other groups. Figure 2 is a graphical representation of the performance of each group as independent classifiers.

Figure 1

Mean AUC for each group

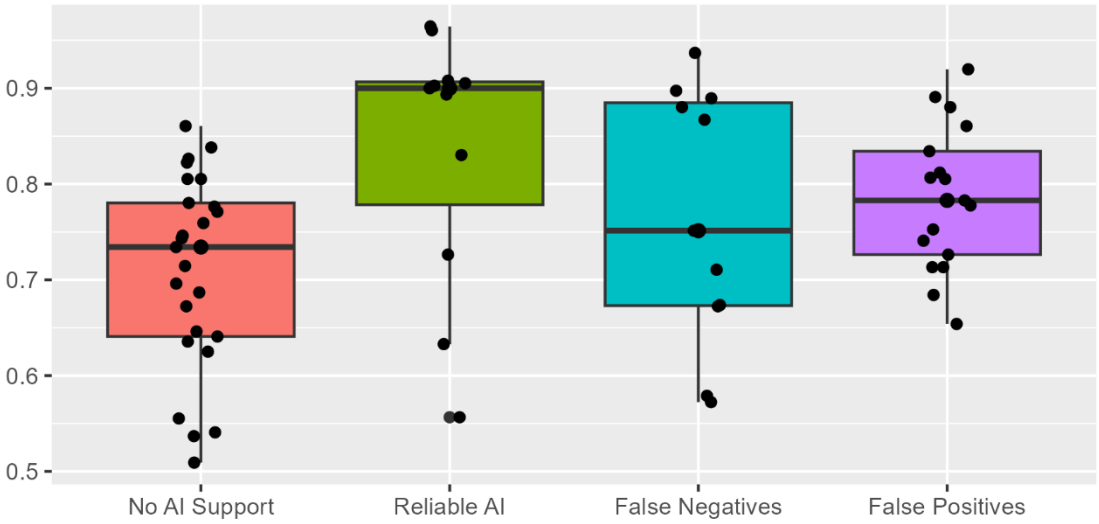
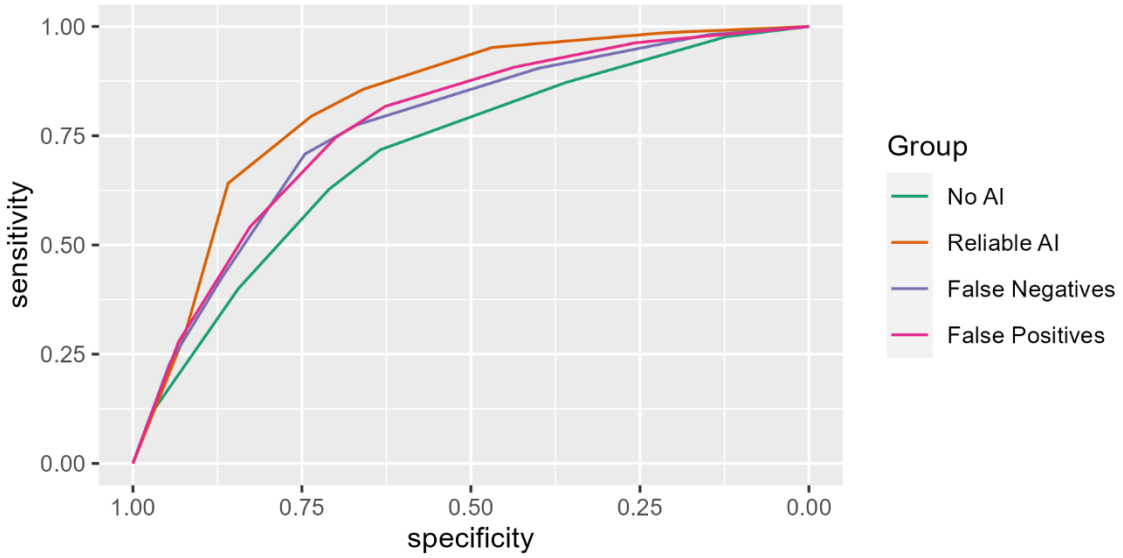


Figure 2

ROC curves for each group



3.1.4 Discussion

The study aimed to examine the effect of using AI compared to no AI support and to assess the potential influence of AI reliability on face-matching performance. It was a preliminary step towards future experiments focusing on trust in AI, as performance was expected to improve if participants trusted and used reliable AI. The study also examined the influence of AI reliability, by comparing the effects of high AI reliability and low AI reliability on performance. In particular, the study provided examined the types of errors that AI can produce and examined the effects of these on performance. Low AI reliability was designed to either make errors where matched identities were labelled as “different”, or mismatched identities were labelled as “same”.

Results confirmed that face-matching performance can be enhanced when given AI support when AI is reliable. This is supported by the finding that the Reliable AI group were better able to discriminate match from mismatch trials compared to the no AI Support group. Results also suggested that using AI support with limited reliability led to an improvement in face-matching performance, however, this improvement was not significant. There was insufficient power to confirm this due to the small and unequal sample size in each group. Previous research has confirmed that users of automated decision aids are sensitive to different levels of reliability but also that individuals tended to disagree with an aid even when it was fully accurate (Wiegmann et al., 2001).

Further insights were gathered from the Low AI Reliability groups. Findings suggested that the specific nature of AI error might not significantly impact face-matching performance. In practice, threshold placement which determines the false rejection and false acceptance rate is often pre-determined (Cavazos et al., 2021). Different systems may have different reliabilities in various face-matching scenarios. Clarifying whether false

alarms and misses have different effects on human performance can help understand the influence of facial recognition technology and may help with threshold placement to optimise performance.

Further research could clarify the impact of low AI reliability in different settings. Research has also shown that motivations induced by monetary incentives can affect face-matching performance (Susa et al., 2019). Examining whether this could produce different results within the context of low AI reliability would be an interesting direction of research. While the current experiment did not find a difference between the impact of the two types of AI errors, it is recognised that the results may be different in real-life settings where the purpose for face-matching, higher stakes and time constraints could be additional factors that influence the significance of AI error types.

This experiment was an initial exploration of the relationship between AI reliability and performance. Whilst the impact of high AI reliability on performance is verified, the experiment has yet to examine whether improvements were related to trust, as trust calibration is theorised to improve performance (Lee & See, 2004). The results are promising in demonstrating that participants made use of the AI. However, the fact that performance was not consistently perfect when using reliable AI indicates that the user did not trust the AI completely. Previous research has confirmed that system reliability is predictive of trust towards automation but does not affect reliance on automation (Chavaillaz et al., 2016). A closer examination of trust in AI would indicate whether participants' performance improved because they trusted and used the AI.

It is recognised that the study assigned different participants to each group. Research has highlighted that there is a large variation in face-matching abilities between individuals (Bindemann et al., 2012), and this individual difference in baseline abilities could

have driven the results rather than AI reliability. This is because different individuals exhibit differences in bias and therefore decision-making in face-matching (Baker et al., 2023). Future research could consider within-participant designs to ensure that individual differences do not impact the face-matching outcomes.

Furthermore, the confidence statements in the current study used a 7-point scale to assess confidence in their face-matching decisions. While this was for calculations of AUC and ROC, there was a possibility of response bias, as participants may have avoided the extreme ends of the scale on difficult trials. Confidence ratings were used as measures of performance, but AUC and ROC do not inform whether changes in performance were a result of changes in sensitivity or bias.

In summary, the current study provided valuable insights into the impact of using AI support in face-matching, showing that performance can be improved when given AI assistance, particularly when the AI is reliable. There are initial findings that AI reliability impacts performance, but further research is required to examine to role of trust.

3.2 Experiment 1

3.2.1 Introduction

Studies on unfamiliar face matching and facial recognition have opened up a new area for research into trust and human-AI interactions. Face-matching refers to comparing two faces presented simultaneously and deciding whether they belong to the same person or different people. Matching unfamiliar faces can be a difficult task. When sorting real-life life photographs into piles based on identity, images of the same person were often perceived differently by unfamiliar viewers while familiar viewers could perform the task relatively accurately (Jenkins et al., 2011). Similarly, pictures of famous faces are processed more accurately compared to unfamiliar faces (Carbon, 2008). Poor accuracy in unfamiliar face matching is not limited to laboratory studies involving naïve participants, as professionals are susceptible to errors too. For instance, passport officers displayed no advantage over the general population on standardised face-matching tasks (White et al., 2014). Error rates are expected to be higher in applied settings where real-life tasks present other challenges, such as an unbalanced ratio of identity matches and mismatches. Mismatches often go undetected in low-prevalence conditions compared to high-prevalence conditions (Papesh & Goldinger, 2014). Matching unfamiliar faces is a difficult task and therefore worthy of further research as it is a widely used form of identity verification.

Currently, facial recognition systems in real-life contexts mostly include a human-in-the-loop, where the human operator is required to review the output of a facial recognition algorithm and make a final decision (Sanchez del Rio et al., 2016). In an identification task using passport photographs, facial reviewers asked to compare an image to a candidate list of possible matches made an error on average in every other trial (White, Dunn, Schmid &

Kemp, 2015). When provided with a target image and asked to compare with a candidate list of images, ordered and ranked by their degree of similarity to the target face image, the length of the candidate list presented to reviewers can significantly affect performance in an unfamiliar face-matching task (Heyer, Semmler & Hendrickson, 2018). This is similar to the findings in research on fingerprint examiners and automated fingerprint identification systems (Dror & Mnookin, 2010). These findings indicate that the way outputs of an automated face-matching system using AI are presented to users requires careful design to reduce bias and to aid the decision-making process in a meaningful way. Understanding the influence of using AI on decision-making is necessary to optimise human-AI interaction.

Bias towards automation in decision-making is referred to as automation bias and is evident in other domains and tasks. For example, interactions between radiologists and computer-aided diagnosis have also been shown to be imperfect leading to suboptimal diagnostic performance (Jorritsma et al., 2015). Research on interactions between healthcare professions and clinical support systems has provided evidence of automation-induced complacency and insufficient monitoring of automation (Goddard et al., 2012). Participants making omission errors defined by the failure to detect mistakes made by decision support systems provides evidence of automation bias (Lyell et al., 2018). Relying on automation when they are imperfect is problematic. Unreliable automation has been shown to negatively affect performances in high workload conditions, despite being aware that imperfections existed (Wickens & Dixon, 2007). Research has also found that participants in an x-ray screening task tended to follow recommendations of automated decision aids rather than using their own abilities regardless of whether recommendations were accurate, in a simulated airport security procedure (Davis et al., 2020). These tasks are conceptually similar to face matching, requiring perceptual and cognitive processes in their

decision-making of a categorical response. Given the similarities with the decision aids discussed, there is a potential connection to the literature on face matching and facial recognition literature.

Effective interactions with technology require trust. Trust is complex but can be broadly categorised into dispositional trust, situational trust, and learned trust (Hoff & Bashir, 2015). Dispositional trust is relatively stable over time, influenced by factors such as age, gender and personality. Independent of the context and type of automated system used, dispositional trust such as indicated by propensity to trust is an individual characteristic that is important in human-AI interactions and can be referred to as the tendency to be trusting of automation in general (Merritt & Ilgen, 2008). To reduce incidences of misuse and automation-induced errors, trust calibration may be necessary. Miscalibration of trust can result in disuse, which refers to the neglect or underutilisation of automation and misuse refers to the over-reliance on automation, (Parasuraman & Riley, 1997). Trust calibration is the process of matching a user's level of trust with the given reliability of the automation (Lee & See, 2004). When trust is calibrated, the human operator can search for alternative resources to support the decision-making process when the AI provides unreliable recommendations. Alternatively, the human operator can trust AI advice when accurate recommendations are provided.

The aim of Experiment 1 was to evaluate the influence of AI reliability on face-matching. In line with previous findings showing that system reliability impacts performance and trust (Chavaillaz et al., 2016), the hypothesis was that low AI reliability would have a negative influence on both face-matching accuracy and trust in the AI. Using a similar setup to the experiment by Fysh and Bindemann (2018), Experiment 1 made use of consistent or

inconsistent pairings of faces and AI labels to vary the reliability of the AI system. It was hypothesised that using a facial recognition algorithm with low reliability that contains inconsistent labels reduces trust in the system, as indicated by lower self-reported trust ratings in the AI. By comparing face-matching performance under conditions of high and low AI reliability with a control group (where no AI support was provided), Experiment 1 aimed to examine whether human participants supported by AI with high reliability would perform better compared to when no AI support or AI with low-reliability support was provided.

3.2.2 Method

Participants.

A total of 110 participants were recruited through *Prolific* (*www.prolific.com*). Participants were adults aged 18-35, self-identified to be British. Participants were compensated using *Prolific's* payment system, at a rate of £7.00 per hour.

Materials.

Glasgow's Face Matching Test (GFMT): The short version of the GFMT consisted of 40 pairs of faces, photographed in a full-face view taken with different cameras with all faces displaying a neutral expression (Burton, White & McNeil, 2010). Images in the GFMT contain only the face. Of the 40 face pairs, 20 depict the same identity. The GFMT was used as a pre-test to compare the baseline performance of participants between experimental groups.

Kent Face Matching Test (KFMT): The short version of the KFMT consists of 40 pairs of faces, each containing one student ID stylse of image and one high-quality portrait taken at least three months apart (Fysh & Bindemann, 2018). Student ID photos are not controlled by expression, pose or image-capture device. Similar to the GFMT, 20 of the 40 face pairs depict the same identity.

Face-recognition: This is a Python package available online under MIT license which uses *dlib's* facial recognition algorithms (version 1.3.0.). All KFMT face pairs used for testing were processed by *face-recognition* and the dissimilarity score for each face pair was obtained. The score is a measure of dissimilarity between the two images, where the lower the score, the higher the similarity between two faces. The algorithm was unable to detect a

face in one image and this image was included in the experiment but excluded from the analysis.

PsychoPy interface: The experiment was designed on *PsychoPy* (Peirce et al., 2019). The layout of a single trial consisted of a face pair image placed to the left of the screen with radio buttons labelled ‘Match’ and ‘Non-match’ to the top right.

Design

The experiment involved two face-matching tests: GFMT and KFMT. Both tests made use of a mixed-participant design, with the variables of trial type (match or non-match) as the within-participant variables and group (High AI Reliability, No AI or Low AI Reliability) as the levels of the between-participant variable.

For the KFMT, high-reliability AI provided accurate results 100% of the time, while low-reliability AI had an error rate of 30% and no AI support was provided in the control group. A label of ‘same’ or ‘different’ was available beneath the face pair images on each trial in conditions where AI support was provided. AI reliability differed between the experimental conditions. In conditions with high AI reliability, labels were fully consistent, reflecting the actual classifications made by *face-recognition,* which were also objectively accurate. In conditions with low AI reliability, some of the labels were inaccurate: 6 match trials (depicting the same identity) were labelled “different” and 6 mismatch trials (depicting different) identities were labelled “same”. Images in the inconsistent trials were selected based on their dissimilarity score obtained from the facial recognition algorithm. Matches with the lowest similarity and mismatches with the highest similarity were inconsistently labelled. A face pair depicting a famous politician was included as an attention check to

ensure that participants were engaged in the experiment and were following instructions correctly.

Procedure

Participants were asked to read the Participant Information Sheet and provide consent before proceeding to the online experiment hosted on *Pavlovia* (*https://pavlovia.org/*). All participants were asked to complete the GFMT. Participants were instructed to decide whether two images were of the same person or different people by pressing 's' for same or 'd' for different on their keyboard.

For the KFMT, participants were assigned to one of the three experimental groups. In conditions where AI support was provided, participants were made aware that they would see a label of 'same' or 'different' that was provided by a face-matching algorithm. Participants in these two groups were advised that the AI output presented might not be accurate. In the control group where no AI support was provided, participants were only asked to examine the faces carefully and judge the face pair as match or non-match.

After clicking 'match' or 'non-match', a rating scale appeared prompting participants to rate their confidence in the decision, from 0 to 100. Anchors in this scale were 'Extremely confident' at 100, 'Very confident' at 75, 'Quite confident' at 25 and 'Not confident' at 0. Next, a confirm button appeared allowing the participant to finish the current trial and proceed to the next. In conditions where AI support was provided, participants were asked to rate their trust in the output given in every trial.

Methods of analysis

Signal detection theory (SDT) can be used to analyse performance on classification tasks (Kostopoulou et al., 2018) and was applied to the face-matching task in Experiment 1.

Participants were expected to make a judgement on whether a face pair is positive (match) or negative (mismatch) which could result in one of four types of responses: Correct responses include identifying a match to be a match (hit), a mismatch to be a non-match (correct rejection); incorrect responses include falsely identifying a match (false alarm) or falsely identifying a mismatch (omission). Experiment 1 used *d*' and *c* as measures of sensitivity and bias, by calculating the difference between the mean hit and false alarm rates and measuring the criterion shift (Stanislaw & Todorov, 1999). The current study used the R package pROC (version 1.18.4) to calculate AUC, which represents the area under the receiver operating characteristic curve (ROC), a measure widely used to assess and compare the performance of classifiers without the implementation of a specific dissimilarity or similarity score threshold.

3.3.3 Results

Of the 110 participants who completed the experiment, 10 participants were removed from the data analysis as they did not meet the study criteria (1 participant), failed to follow the instructions correctly or did not pass the famous face trial (9 participants). We ended up with 32 participants in the No AI group, 33 participants in the High AI Reliability group and 35 participants in the Low AI Reliability group.

GFMT Performance.

The GFMT was included as a pre-test to verify that participants performed similarly across groups before completing the KFMT. Percentage accuracy for each group was obtained and calculated by the number of correct trials divided by the total number of match or non-match trials. To examine differences in sensitivity and bias, *d*' and *c* were also calculated, using the R package *Psycho* (version 0.6.1). Table 2 summarises the descriptive results. ANOVA tests on *d*' [$F(2, 97) = 0.373$, $p = .690$], and *c* [$F(2, 97) = 0.774$, $p = .464$] showed no significant differences between the groups.

Table 2

Percentage accuracy, averaged d' and c in the GFMT

Group	*d'*	*c*	Non-match	Match
	M (SD)	M (SD)	%	%
High AI Reliability	1.57 (0.77)	-0.03 (0.38)	75.45	78.2
No AI	1.74 (0.90)	0.05 (0.44)	79.22	77.0
Low AI Reliability	1.61 (0.76)	-0.07 (0.40)	75.43	79.0

KFMT Performance.

Performance across the groups in the KFMT was also examined, using measures of sensitivity and bias. Table 2 is a summary of the percentage accuracy and the calculations of *d*' and *c* for each group in the KFMT.

Table 3

Percentage accuracy, averaged d' and c in the KFMT

Group	*d'*	*c*	Non-match	Match
	M (SD)	M (SD)	%	%
High AI Reliability	0.84 (0.56)	-0.18 (0.44)	58.94	70.76
No AI	0.86 (0.58)	-0.01 (0.40)	64.38	66.72
Low AI Reliability	0.83 (0.52)	-0.26 (0.43)	56.00	74.29

Results also indicated that there were no significant differences across the groups in average *d*', [F(2, 97) = 0.026, *p* = .974]. Figure 3 shows the distribution of *d*' in each group. However, there was a significant difference between the groups in measures of bias [*F*(2, 97) = 3.102, *p* = .049]. Results of the post hoc test showed that *c* in the Low AI Reliability group was significantly lower than in the control group, indicating a higher tendency to respond 'match' than 'non-match', with an average difference of 0.25 (*p* = .041), compared to when given no AI support. Figure 4 displays the distribution of *c* in each group, showing the increased bias in the Low AI Reliability group.

Figure 3

Distribution of d' by group in the KFMT

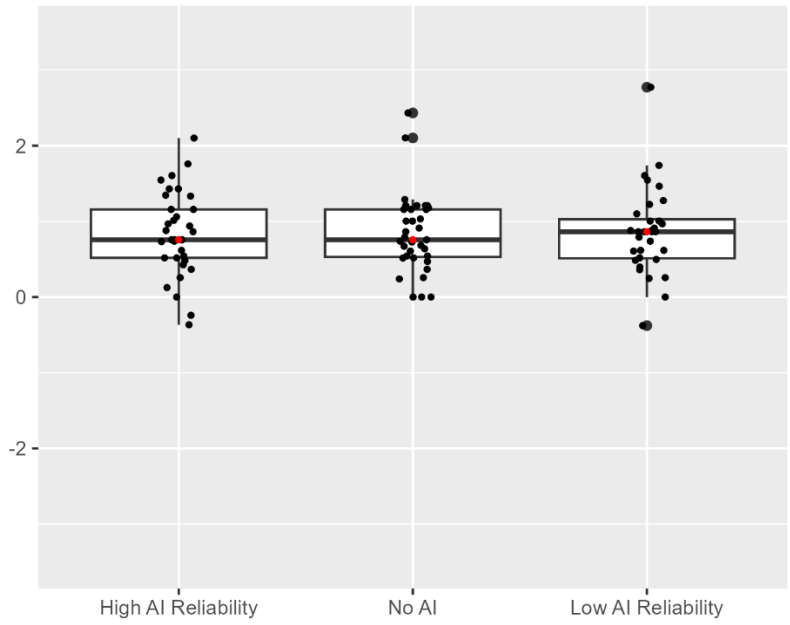
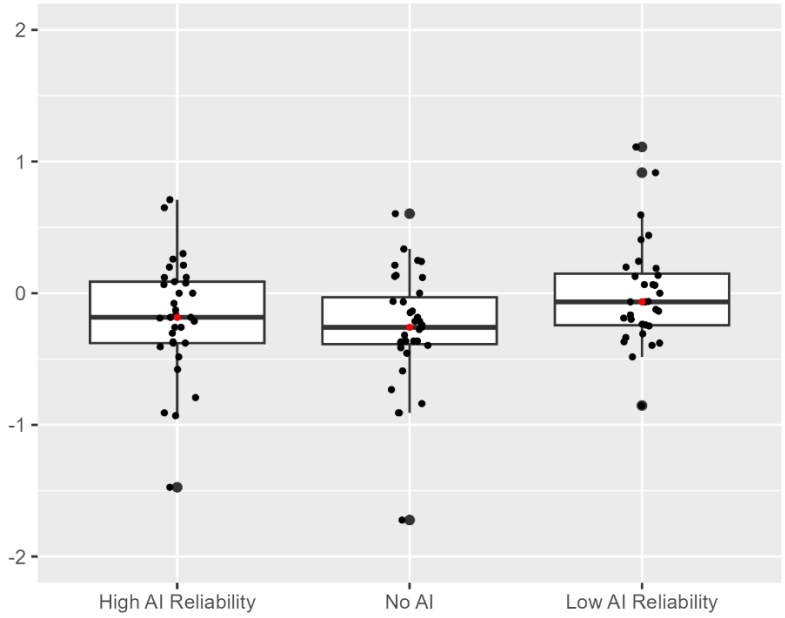


Figure 4

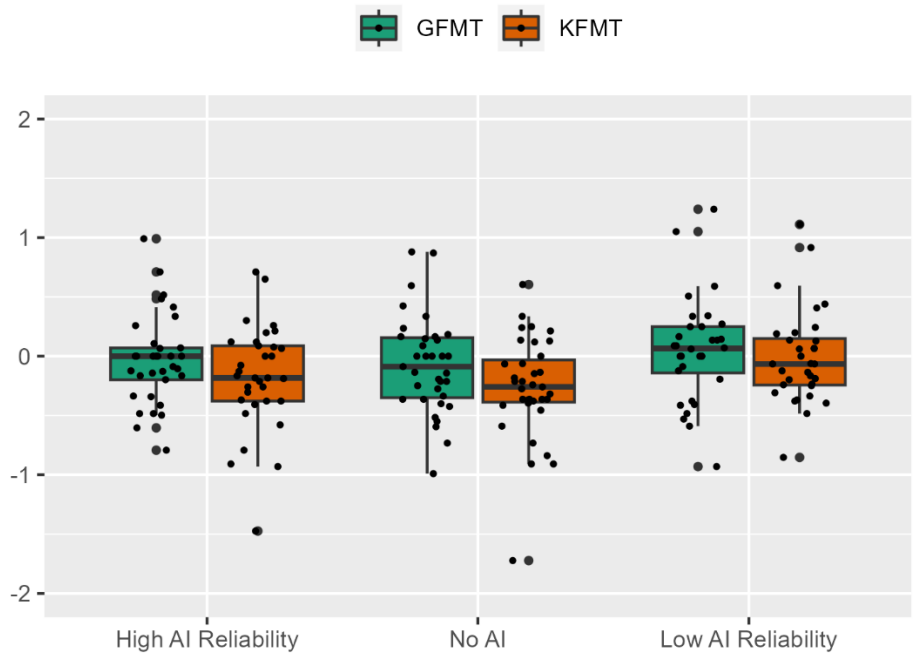
Distribution of c by group in the KFMT



There were no significant differences between the No AI Support (control) group and the High AI Reliability group. However, a more liberal response was also observed in the High AI Reliability group. Further analysis was carried out to compare *c* between GFMT and KFMT. Results of an ANOVA indicated a significant main effect of Test [$F(1, 194) = 5.115$, $p = .025$] and Group [$F(2, 194) = 3.528$, $p = .031$]. Results confirmed the finding of an elevated liberal response from GFMT to KFMT in the Low AI reliability group ($p = .056$). Figure 5 is an illustration of c in both tests

Figure 5

Comparison of c between GFMT and KFMT



Decision Outcomes in the KFMT

Responses on a given trial were categorised as either correct or incorrect, which was referred to as the outcome of the decision. To examine differences in trust, as predicted by group and the type of trial, the following model was used:

Outcome ~ Group + (1 | Participant) + (1 | Trials)

Results confirm no significant differences between the groups (*Est.* = -0.04, SE = 0.07, *Wald Chi-Square* (1) = 0.33, *p* = 0.567), suggesting that AI reliability does not influence the accuracy of a decision.

Trust Rating.

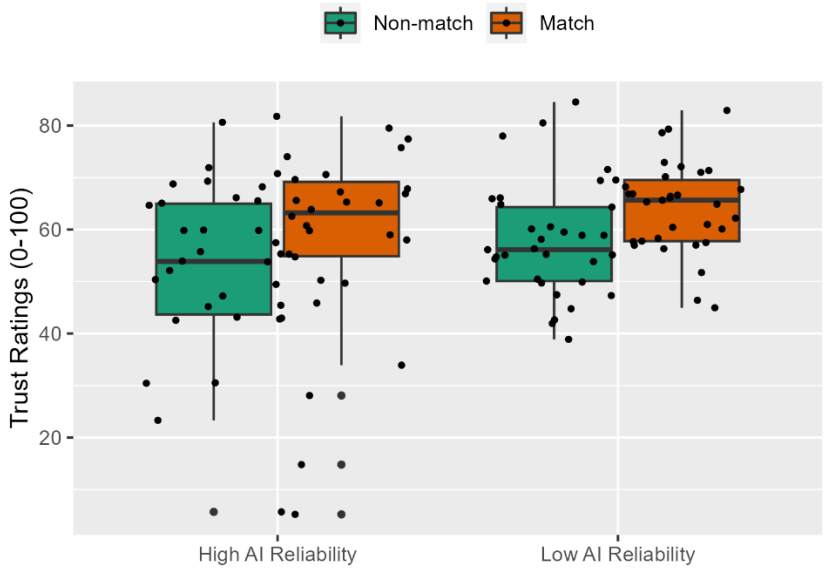
Trust ratings for the algorithm advice were higher on match trials than mismatch trials in both the high AI reliability group (M = 58.48, SD = 8.57) and the low AI reliability group (M = 64.26, SD = 7.03). Trust ratings on mismatch trials in the high AI reliability group were lower (M = 52.39, SD = 8.57) than in the low AI reliability group (M = 56.90, SD = 8.32). The following model was specified to examine differences in trust, as predicted by group and the type of trial:

*Trust ~ Group * Type + (1 + Type | Participant) + (1 | Trials)*

Results indicated no significant main effect of Group groups (*Est.* = -4.452, SE = 3.594, *Wald Chi-Square* (1) = 1.534, *p* = 0.215), or Type (*Est.* = 7.534, SE = 6.072, *Wald Chi-Square* (1) = 1.540, *p* = 0.215). The Group X Type interaction was also not significant (*Est.* = -0.808, SE = 3.210, *Wald Chi-Square* (1) = 0.063, *p* = 0.801). Figure 12 is an illustration of the percentage accuracy in each condition.

Figure 6

Average of Trust Ratings in each group



Fusion of Human Rating and AI Scores.

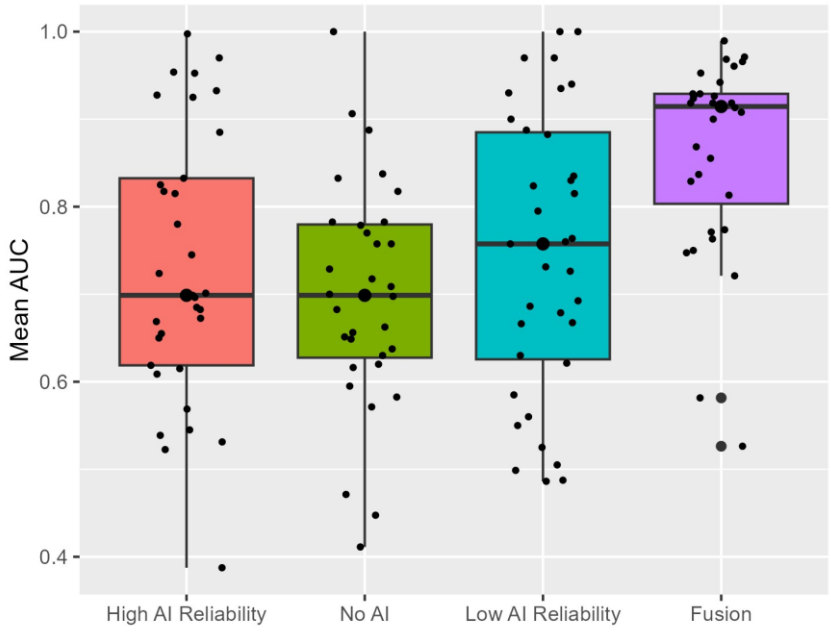
Previous research has suggested that the fusion of human rating and AI scores can improve face-matching performance (Phillips et al., 2018). To verify whether using reliable AI further improves accuracy, an exploratory analysis was carried out to replicate findings on the fusion of ratings. Algorithm scores were rescaled to the range of human ratings and for each face pair, the human rating and algorithm score were averaged and used to calculate an AUC score for each participant.

For each image pair, the algorithm returned a score that represented the dissimilarity of the two faces. The scores were scaled to within the range of all human ratings and averaged with each human participant in the control group to give a fused human algorithm score. These fused scores were used to calculate an AUC for each participant which was then averaged by group, see Figure 7.

The mean AUC score was the highest for the fusion group (M = 0.86, SD = 0.11), followed by the Low AI Reliability group (M = 0.74, SD = 0.16), High AI Reliability Group (M = 0.73, SD = 0.15) and the No AI group (M = 0.70, SD = 0.13). Results of an ANOVA on the averaged AUC show a significant effect of group, [$F(3, 128) = 8.266$, $p < .001$]. Further analysis indicated an average difference of 0.16 between the No AI group and Fusion group in AUC score ($p < .001$), 0.13 for the High AI Reliability group ($p = .001$) and 0.11 for the Low AI Reliability group ($p = .005$).

Figure 7

Boxplots of AUC scores averaged across participants in each group



3.3.4 Discussion

The purpose of Experiment 1 was to gain a better understanding of face-matching performance in human participants when given AI support of high or low reliability, compared to when given no AI support. In particular, the experiment was designed to add to previous findings by examining the influence of AI reliability on trust in facial recognition outputs. This led to the hypothesis that using AI with high reliability that provides consistent information improves face-matching accuracy and using AI with low reliability that provides both consistent and inconsistent information has the opposite effect, compared to using no AI. With links to the literature on trust and automation, it was hypothesised that differences in performance between the groups could be reflected in self-reported trust ratings. Previous research has found trust to be a factor that is important in all human-automation interactions. Thus, trust was also expected to be important in the interactions between humans and facial recognition algorithms in making face-matching decisions.

Results showed no significant differences in sensitivity between conditions with high or low AI reliability and no AI support. This appears to suggest no obvious effects of using AI with high or low reliability on sensitivity as neither group performed significantly better than the control group where no AI support was provided. This could mean that AI does not impact overall face-matching performance. This finding is further confirmed by analysing the correctness of a decision predicted by participant group, reinforcing that AI reliability had no influence on the accuracy of face-matching decisions.

Despite not finding an effect of using AI or AI reliability on sensitivity, results showed a significant shift in response bias. Findings suggested that AI with low reliability increased the likelihood of responding ‘match’, even on non-match trials. A shift in bias is also evident when results obtained with the experimental KFMT stimuli were compared to earlier results

obtained with the preliminary GFMT stimuli. This perhaps suggests that using AI in face-matching introduces bias, and this effect is more pronounced when given AI with low reliability. This could mean that in applied settings, false alarm rates of the human operator may increase when face-matching with AI.

Results also indicated that trust was generally higher on match trials than on non-match trials. However, the analyse revealed that AI reliability was not a significant predictor of trust. Previous research examining human-automation interactions has demonstrated that trust decreases after seeing credible systems make errors (Madhavan & Wiegmann, 2005). Current findings would suggest that trust was not sensitive to the performance of the AI.

Results from the fusion of confidence ratings provide additional support that humans and AI further improve accuracy when their responses are combined together. This is consistent with previous research that the wisdom-of-crowds effect (Surowiecki, 2004) can be applied to face recognition (Phillips et al., 2018). The findings of the current experiment indicate that independently fusing algorithm scores and human ratings produces more accurate decisions, by improving the performance of the human. It is also important to highlight that the algorithm did best on its own, suggesting that the human is the limiting factor in this interaction. This is in line with previous research demonstrating that using accurate automated facial recognition systems can improve performance but face-matching with AI support often fails to reach the level of accuracy AI systems can achieve alone (Carragher, 2023). This finding highlight that the partnership between humans and AI should be further explored to optimise the interaction as in cases where AI fails, human oversight is still essential.

Experiment 1 made use of a between-participant design, and it is possible that different participants could have employed different strategies in their uses of the labels which led to insignificant differences in sensitivity, as participants may have chosen to use or disuse the algorithm. While the results of the GFMT indicate no significant difference between the groups prior to completing the KFMT test, the way each individual made use of the AI was not clear. Differences in trust ratings between the groups could also be explained by differences in individuals' propensity to trust, a variable that was not examined in the current experiment.

There are limitations to the experiment that must be acknowledged. For instance, there could a better control of other influential factors by matching participants on their demographics. The current study involved participants within a similar age range and all recruited online. However, better matching of other backgrounds such as occupations and experience is ideal. For example, experience prior training may enhance abilities in face matching in forensic examiners (White, Phillips, Hahn, Hill & O'Toole, 2015). Familiarity with the KFMT is another uncontrolled variable that may have influenced the results unfavourably, as repeated exposures to the same faces add to the familiarity-based advantage (Clutterbuck & Johnston, 2005) and participants may have seen the same images in other experiments. If possible, further research is recommended to adopt a within-participant design so that individuals can act as their own control or assess familiarity with the test material with post-test questions.

Like all latent constructs, trust is difficult to measure. A self-reported measure of trust has the limitation that participants may not be honest with their responses and fail to consider their implicit attitude toward automation which may have influenced trust (Merritt

et al., 2013). As trust is conceptualised differently in different fields and disciplines, different aspects of trust can lead to different ways of measuring trust (Lewicki et al., 2006). For example, measurements of the propensity to trust automation surveys are more applicable when trust is conceptualised as an attitude or intention (Jessup et al., 2019). Sources of variability in human-automation trust include the human operator, the environment, and the automated system (Hoff & Bashir, 2015), which adds to the difficulty in measuring trust. Future studies could focus on a specific layer of trust, such as dispositional trust, situational trust or learned trust (Marsh & Dibben, 2003), or use a variety of trust measures to better capture changes in trust that are occurring throughout the interaction. Different measures of the same construct would strengthen the findings.

The relationship between trust and behaviour can be better defined and examined in future studies. Research on decision support tools has distinguished between reliance and compliance as behaviours present in human interactions with imperfect automation (Meyer et al., 2014). Compliance occurs when the human operator obeys when the automation gives a piece of incorrect advice and reliance is when the operator fails to detect an error when not alerted. The framing of the reliability of the automation is important in its utilisation by operators (Lacson et al., 2005), which could be more carefully considered in further research.

Conclusion

In summary, Experiment 1 found an effect of AI on response bias, showing that participants had the tendency to respond match when using AI with low reliability in particular. The implication of this in applied settings is that using AI increases false alarm rates in humans. Despite having an impact on bias, there was no effect on sensitivity.

Experiment 1 also confirmed that fusing human ratings and AI scores improved the face-matching accuracy of the human, but not the AI, as AI on its own achieved the highest accuracy. Possible individual differences in approaches to the task and perceptions of the AI are alternative explanations to the results. Future studies could further verify the involvement of trust in using facial recognition algorithms, and continue to explore the calibration of trust as a way to facilitate the interaction between humans and AI.

Chapter 4: AI Transparency and Dissimilarity Scores

4.1 Pilot Study 2

4.1.1 Introduction

Face-matching involves comparing two faces simultaneously and deciding whether they have the same or different identities. It is an important task that can be supported by facial recognition technology, for example, at border control. Travellers who cannot verify their identity through facial recognition systems are transferred to human officers to perform a manual identity check (Sanchez del Rio et al., 2016). However, research has suggested this sequential setup of the AI-human face-matching process to be imperfect, as human operators often get biased by the outputs of algorithms.

When face pairs are accompanied by inconsistent labels such as ‘same’ for mismatched face pairs and ‘different’ for matched face pairs, the face-matching accuracy of the decision-maker decreases (Howard et al., 2020). The ability to work cooperatively with AI systems is important as the consequences of errors could include wrongful convictions or security breaches. In addition to the demands and difficulty of matching unfamiliar faces, human operators have to adapt to work with AI systems in applied settings. Trust is a concept that has not been explored in the context of face-matching and is hypothesized to be of significant influence as it is an important aspect to consider in human-AI interactions. Trust calibration, the process by which operators learn to adjust their trust towards automation based on their actual performance and capabilities, could offer a path to improving face-matching accuracy (McGuirl & Sarter, 2006). AI that can be easily understood and analyzed by humans are considered transparent or interpretable and this appears to address the black box issue (Hagras, 2018). Providing explanations for AI outputs

helps calibrate trust (Zhang et al., 2020). The current experiment serves as a pilot study with the eventual aim of examining dissimilarity scores, a measure of the similarity between two faces, as a way to add transparency to a facial recognition system.

Currently, automated facial recognition systems function under the supervision of human operators who intervene via a computer interface when the system is unable to resolve an identity verification (Gaves et al., 2011). Applying the wisdom of the crowd effect (Suroweicki, 2004) to face-matching, previous research has shown that fusing human ratings and normalised AI scores improves face-matching accuracy (O'Toole et al., 2007). Team performance can be further improved by fusing algorithm outputs with ratings made by professionals such as forensic facial examiners (Phillips et al., 2018). However, fusing scores made independently by humans and AI is different to monitoring and validating prior judgements made by facial recognition technology. When face images are paired with inconsistent information, such as labels suggesting 'same' on trials containing face pairs depicting two different identities or 'different' on trials containing face pairs of the same identity, face matching accuracy decreases (Fysh & Bindemann, 2018). Further research has confirmed these findings, demonstrating that inconsistent labels shift participants' internal criteria used in face-matching judgements (Howard et al., 2020). Facial recognition technology in applied settings introduces errors, despite that in theory, human-AI teams should increase general accuracy.

Making mistakes by following incorrect advice or failing to act when not prompted to do so by decision support systems are examples of automation-induced errors (Parasuraman & Manzey, 2010). The errors observed could be related to issues of trust. When human operators under or overtrust automated decision aids, they often underutilise

or overly rely on the decision support system (Parasuraman & Riley, 1997). Trust determines the willingness of human operators to rely on automation, and sources of variability can be categorised into dispositional trust, situational trust, and learned trust (Hoff & Bashir, 2015b). Trust calibration, the process of matching a user's trust in the AI and the AI's actual capabilities, is suggested to have positive effects on human-AI interaction (Lee & See, 2004).

However, using imperfect AI systems can have detrimental effects on trust. After observing AI make errors, participants often distrust reliable aids unless additional explanations are provided on why errors might occur (Dzindolet et al., 2003). It is evident that system reliability is predictive of performance and trust (Chavaillaz et al., 2019). Trust has not been explored in the context of face-matching, despite being an important influence in human-AI interactions. Trust can increase the use of an automated system (Khastgir et al., 2017), and calibrated trust can improve performance.

Presenting confidence information on a system's ability to perform a given task, aids in trust calibration and reduces errors in decision-making (McGuirl & Sarter, 2006). Providing an explanation for an AI output or simply adding more information can aid trust calibration (Hussein et al., 2020), as well as displaying a confidence score for a given AI prediction (Zhang et al., 2020), and in this study, it was hypothesized that dissimilarity scores can work the same way.

To explore the idea of whether presenting dissimilarity scores can be a tool to enhance trust calibration when given an imperfect AI, the pilot study aims to first verify the influence of presenting dissimilarity scores, ascertaining whether these are even used in the decision-making process.

Having an accurate perception of a system's actual level of reliability is predictive of performance (Merritt et al., 2015b). Dissimilarity scores provide a quantitative measure of difference and may offer valuable guidance as individuals can compare between face pairs. Measures of facial similarity such as Euclidean distance are closely related to participant ratings of face similarity and are suitable to be used as perceived similarity (Tredoux, 2002). The pilot study aimed to understand how face-matching performance is influenced by AI support, particularly in the form of dissimilarity scores. Whether and how AI support in this format is used in face-matching is uncertain and would be useful for further experiments investigating trust calibration. It was hypothesized that presenting dissimilarity scores in face-matching would lead to improved performance. This will be demonstrated by higher percentage accuracy and measures of performance in sensitivity and bias.

4.1.2 Methods

Participants.

A total of 32 volunteers were recruited via the University of Glasgow subject pool. Participants were asked to read the Participant Information Sheet and provide consent before proceeding to the online experiment hosted on *Pavlovia* (*https://pavlovia.org/*). The inclusion criteria were that participants must be students at the University of Glasgow participating for credits for their course. There was no monetary compensation but participant was given course credits for their participation. Participants with known prosopagnosia were not eligible to participate in the face-matching study as stated in the advert.

Materials.

Kent Face Matching Test (KFMT): The short version of the KFMT consists of 40 pairs of faces, each containing one student ID style of image and one high-quality portrait taken at least three months apart (Fysh & Bindemann, 2018). Student ID photos are not controlled by expression, pose or image-capture device.

The Python library *Deepface*, available online under the MIT license, was used to process images taken from the KMFT. Using the *FaceNet* model (Schroff, Kalenichenko & Philbin, 2015), a list of dissimilarity scores for each pair of images from the short version of the KMFT was obtained. This model is documented in the literature and benchmarked against the LFW dataset (Huang et al., 2007).

PsychoPy interface: The experiment was designed on PsychoPy (Peirce et al., 2019). The layout of a single trial consisted of a face pair image placed at the centre of the screen with its dissimilarity score placed at the bottom right of the face pair.

Design.

The experiment used a between-participant design and compared the effects of using AI support with a control group on face-matching performance using images from the KFMT

Procedure.

Participants were asked to read the Participant Information Sheet and provide consent before proceeding to the online experiment hosted on *Pavlovia* (*https://pavlovia.org/*). All participants were asked to complete the KFMT. Participants were instructed to decide whether two images were of the same person or different people by pressing ‘s’ for same or ‘d’ for different on their keyboard.

Participants were assigned to one of two conditions and completed the task either with or without the presence of dissimilarity scores. All participants were encouraged to answer as accurately as they could. Participants who were provided with dissimilarity scores were explained how to use the dissimilarity scores to make a decision.

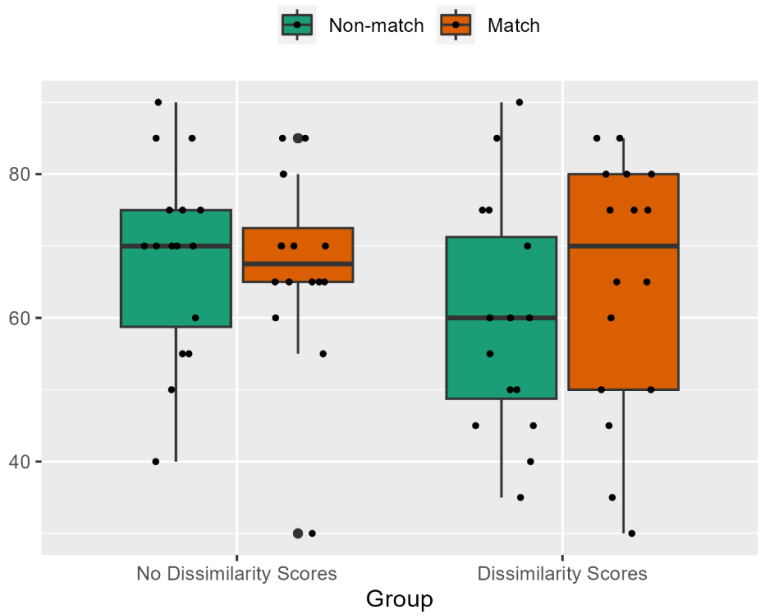
4.1.3 Results

The control group that was not provided with algorithm support had a mean percentage correct of 68.43% and an SD of 13.51 on mismatches and 67.50% on match trials with an SD of 13.17. The group who was given dissimilarity scores along with the face image pairs had a mean percentage correct of 59.69% for mismatches and SD of 5.86 and 64.69% for matches with SD of 17.84.

A two-way ANOVA was conducted to examine the influence of trial type and group on percentage accuracy. Results showed that there was no significant interaction ($F(1,60) = 0.609$, $p = .438$), or main effect of group ($F(1,60) = 2.311$, $p = .134$) or type ($F(1,60) = 0.285$, $p = .595$). Figure 8 is an illustration of the percentage accuracy in each group.

Figure 8

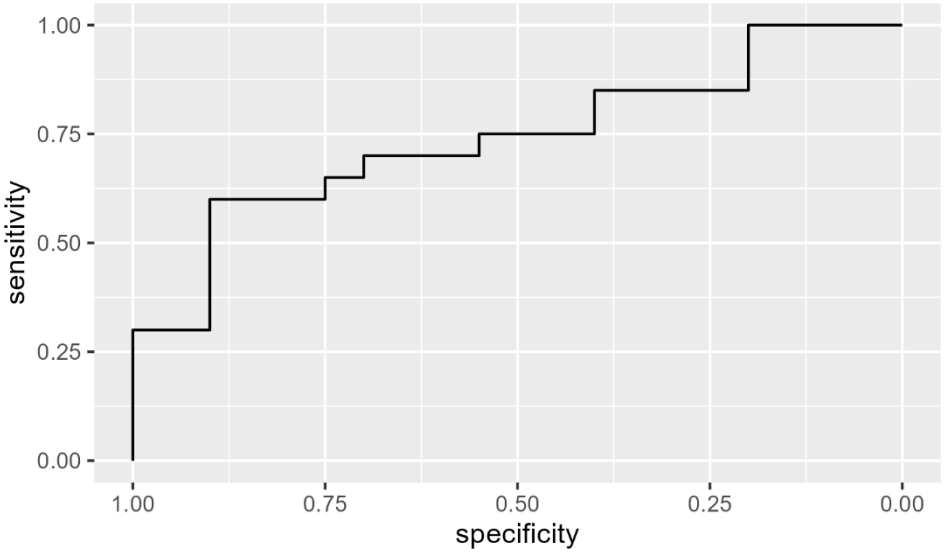
Percentage accuracy for each group in the KFMT



The ROC curve for the algorithm model was computed using the R package *pROC* and the AUC for the model used on the KFMT images was 0.74. Figure 9 is a plot of the ROC curve.

Figure 9

ROC curve



Results also indicated that there were significant differences between the two groups in average d', [F(1, 62) = 5.382, *p* = 0.024]. However, there was not a significant difference between the groups in measures of bias, in c [F(2, 62) = 1.055, *p* = .308]. Figure 10 and Figure 11 are plots of *d'* and *c* for each group, measuring sensitivity and bias.

Figure 10

Distribution of d' by group in the KFMT

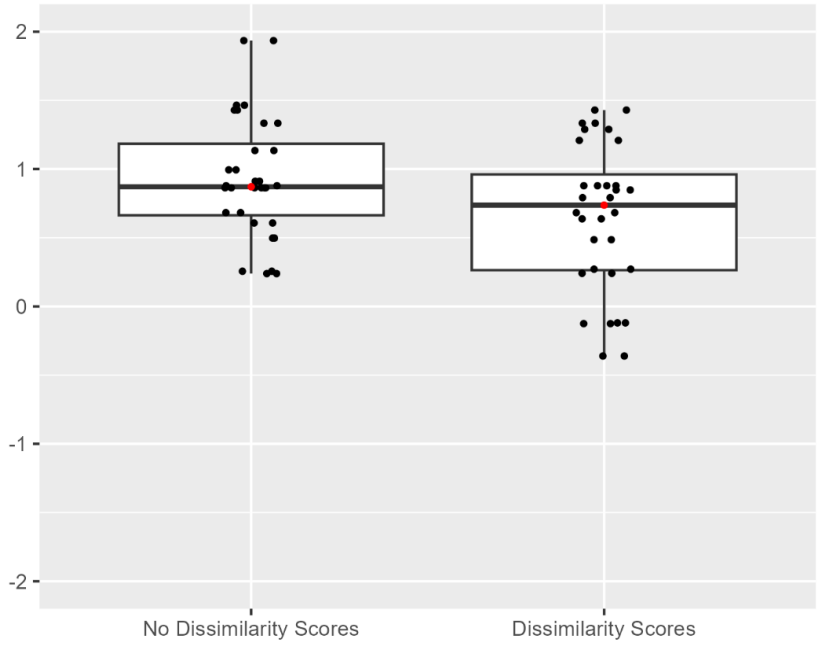
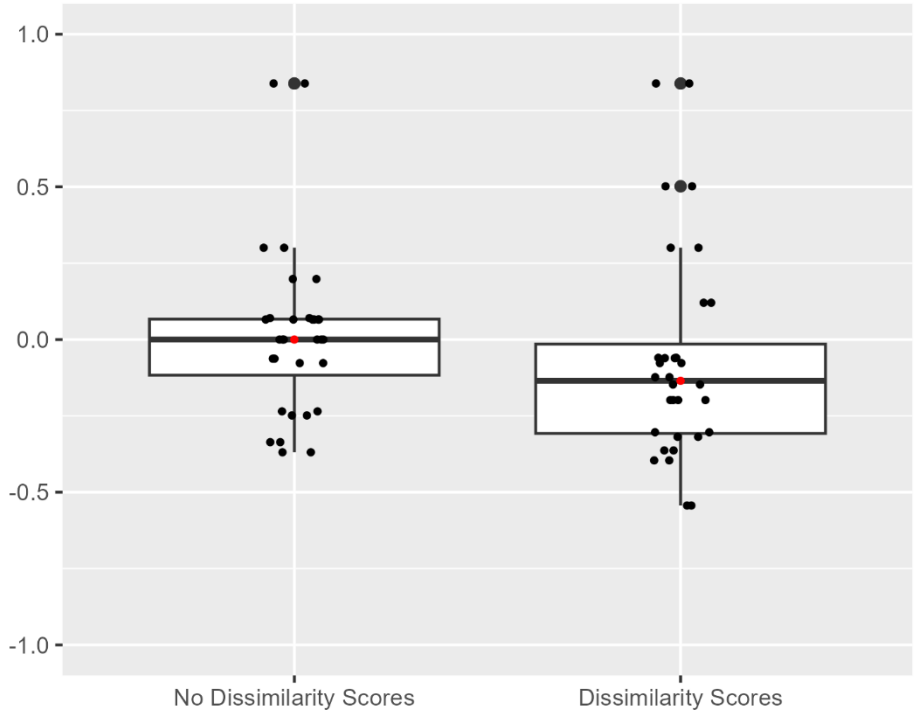


Figure 11

Distribution of c by group in the KFMT



4.1.4 Discussion

The purpose of the pilot study was to gain a better understanding of the effect of using dissimilarity scores in face-matching performance. The current pilot study examined face-matching performance in participants when given AI support in the form of dissimilarity scores, compared to participants who were not given AI support. It was expected that algorithm scores could positively influence face-matching decisions by providing a graded measure of dissimilarity.

In contrast to the hypothesis, results on percentage accuracy suggested that simply presenting algorithm scores may not necessarily be beneficial on face matching performance. Improvement in face-matching might be specific to fusing independent human ratings with algorithm scores (O'Toole et al., 2007). Results therefore suggest that

presenting dissimilarity scores alone might not be sufficient in improving percentage accuracy.

An alternative explanation is that other variables have a more significant impact on human performance, such as the performance of the AI. If the algorithm used in this study had been more accurate, improvements might have been more evident. This is in line with research demonstrating that system reliability affects human performance in high workload conditions (Wickens & Dixon, 2007). As the reliability of the AI was not perfect, participants may have found the AI scores rather ambiguous.

This is further supported by results on sensitivity, showing that participants' ability to distinguish between a match and a non-match face-pair reduced when given AI support. There was no significant impact on bias suggesting that participants' tendency to respond match or non-match was similar in both groups. The current pilot demonstrated the interaction between human and a genuine but imperfect AI system. This serves as a foundation for future experiments investigating whether trust in AI is the mechanism underlying the observed.

Limitations

Current findings did not find statistically significant results in terms of percentage accuracy. There are several explanations that could account for this finding. One potential explanation is that participants may not have fully understood how to best make use of the dissimilarity scores. In the study, participants only received brief instructions, which may not have been sufficient. Future studies could provide more comprehensive training or conduct checks to ensure that all participants have the same, and clear understanding AI dissimilarity scores.

Related to this, the format in which information is presented may also have been important. Information presented in bar graphs may be processed quicker than tables (Brewer et al., 2012). The information could have been useful in the decision-making process, but the use of numerical values may have been processed ineffectively. The format in which AI outputs are conveyed to human operators should be considered in future research.

It is also important to acknowledge that the sample size in this study was relatively small, and primarily consisted of university students. It is recognised that further research with large and more diverse samples might produce different outcomes.

To conclude, this pilot study aimed to explore the potential benefits of using dissimilarity scores generated by AI to enhance face-matching accuracy. While the findings did not reveal statistically significant results regarding percentage accuracy, several important insights were gained.

A limitation of the pilot study was that participants' understanding of how to effectively use dissimilarity scores was assumed rather than verified. Thus, it remains unclear whether changes in performance were related to the reliability of the AI, or the specific format that dissimilarity scores were presented in. Previous research examining the effects of transparency of classifier systems on performance found no difference between different formats of presenting confidence information (Ingram et al., 2021). Future research could benefit from providing comprehensive training and conducting checks to ensure participants had a clear understanding of dissimilarity scores and how they can be used in the face-matching process.

While this pilot study did not demonstrate the immediate impact of AI-generated dissimilarity scores, it serves as a valuable foundation for future research examining the possibility of trust calibration. Understanding the effects of AI support and the presentation format of AI outputs is useful in improving the effectiveness of AI assistance in decision-making tasks and understanding of human-AI interactions.

4.2 Experiment 2

4.2.1 Introduction

Artificial Intelligence (AI) appears to help address the long-studied problem of unfamiliar face matching in humans, which has previously been shown to be highly error-prone and processed differently compared to familiar faces (Megreya & Burton, 2006). Research has further indicated that fusing the responses of the best-performing group of human face specialists with the highest-performing AI produces the most accurate face identification results (Phillips et al., 2018). Collaboration between humans and technology is not perfect, however, and errors occur when the operator misuses automation by overtrusting or disuses automation as a result of under-trusting (Parasuraman & Riley, 1997). The current study aims to examine the concept of trust in automation in the context of face-matching to further understand the interaction between humans and facial recognition algorithms.

Unfamiliar face-matching is required in many applied settings, for example, for identification purposes in security situations where the process of unfamiliar face-matching has increasingly been aided by facial recognition technology. Face recognition is widely accepted as a method of biometric identification, with an algorithm accuracy of approximately 92% (Agrawal & Singh, 2015). Facial recognition technology makes use of state-of-the-art algorithms and has been shown to outperform observers in tests comparing images that are considered to be easy or of moderate difficulty (O'Toole et al., 2012). However, AI performance is comparable to humans under conditions considered to be challenging by algorithm (Phillips & O'Toole, 2014), and its performance is surpassed by forensic facial identification examiners (White, Jonathon Phillips, et al., 2015). Currently,

humans still monitor and supervise facial recognition technology and are expected to continue to interact with such technology in the near future. Therefore, research exploring ways to facilitate human-AI interaction in this domain has practical significance.

Despite the advancement of facial recognition algorithms, face matching appears to be susceptible to different types of errors induced by automation. In an identification task using real-life passport photographs, facial reviewers asked to compare an image to a candidate list of possible matches made an error on average in 1 of every 2 candidate lists (White, Dunn, Schmid & Kemp, 2015). When provided with a target image, and asked to compare with a candidate list of images, ordered and ranked by their degree of similarity, the number of candidate matches presented to reviewers can significantly affect performance in an unfamiliar face-matching task (Heyer, Semmler & Hendrickson, 2018). This is similar to the findings in research on fingerprint examiners and automated fingerprint identification systems (Dror & Mnookin, 2010). The user interface for automated systems, therefore, requires careful design to reduce bias and to aid the decision-making process in a meaningful way.

Mimicking face-matching tasks in a security context at passport control, a research study has shown that decisions such as those made by automated face recognition software impacted performance and demonstrated that face pairs inconsistently labelled as ‘same’ or ‘different’ reduced accuracy in face matching (Fysh & Bindemann, 2018). This demonstrates that text cues may be detrimental to performance when they are inaccurate. In other words, when face pairs were inconsistently labelled as ‘different’ on match trials, or ‘same’ on mismatch trials, face-matching accuracy reduced. Likewise, another study has shown a similar bias as participants were mostly correct when no prior information was given but

introducing labels biased their certainty judgements (Howard, Rabbitt & Sirotin, 2020). The current study will use a similar setup by examining the difference in accuracy between consistently and inconsistently labelled trials.

Trust can be referred to as a willingness to accept vulnerability (Mayer et al., 1995). In the context of automation and AI, trust can be defined as an individual's attitude towards an (automated) agent being helpful in achieving the individual's goals in situations characterised by uncertainty and vulnerability (Lee & See, 2004). The development of trust in humans and in automation is comparable, but differences in reactions to automated or human advice exist (Madhavan & Wiegmann, 200b), as human operators are more sensitive to the errors made by technology than human advisors. Despite having a natural propensity to trust machines, the development of trust in a system is dependent on its reliability. For instance, research has indicated that users of automation tend to rate decision support aid as less trustworthy and reliable after observing errors being made (Dzindolet et al., 2003). Trust is an important factor that mediates the interaction between human AI and has not been explored in previous studies on face matching and facial recognition technology.

Trust can be broadly categorised into dispositional trust, situational trust, and learned trust (Hoff & Bashir, 2015). Dispositional trust is trust that is relatively stable over time, influenced by factors such as age, gender and personality. Independent of the context and type of automation, propensity to trust is an individual characteristic that is important in human-AI interactions. Propensity to trust machines is the tendency to trust automation in general (Merritt & Ilgen, 2008). Propensity to trust can be measured using surveys and can be adapted to be context-specific to be more reliable and predictive of behavioural trust (Jessup et al., 2019). Given that each layer of trust is influenced by distinct factors, the

current study will distinguish between propensity to trust and dynamic trust that varies throughout an interaction to understand the influence of trust on behaviour.

Trust affects whether and how an automated decision aid is used, and to reduce incidences of misuse and automation-induced errors trust calibration may be necessary. Trust calibration is the process of matching a user's level of trust with the given reliability of the automation (Lee & See, 2004a). Having an accurate perception of a system's actual level of reliability is predictive of performance (Merritt et al., 2015b). Presenting dynamic information regarding a system's confidence in its ability to perform a task has been shown to improve a user's calibration of trust in an automated decision aid (McGuirl & Sarter, 2006). System transparency appears to be beneficial for the trust calibration process (Yang et al., 2017). For instance, by providing an explanation for its action or simply adding more information (Hussein et al., 2020). Accompanying AI predictions with a confidence score also appeared to aid trust calibration (Zhang et al., 2020). In a similar way, the current study aimed to explore whether providing additional AI information can help calibrate trust.

Facial recognition systems are susceptible to the black box problem. AI can involve a system of deep neural networks containing operations and components that are largely hidden from the user and are often referred to as black box models, which creates both practical and ethical issues regarding the applications of the system (Guidotti et al., 2018). In particular, the black box problem raises concerns about the trustworthiness of AI systems and improving the transparency of the system is a path to improving trust (von Eschenbach, 2021). For facial recognition algorithms, problems arise when the data used to train the algorithm is not transparent, leading to demographic biases in performance and accuracy. For instance, reports suggest that facial recognition algorithms produce more errors

matching people of a certain race (Grother, 2019). By understanding how an AI reaches a decision, human users can adopt behaviours to counteract the algorithmic bias. Explainable AI provides visibility into the process behind AI decisions and predictions and has been proposed to unmask black-box models (Rai, 2020). In general, system transparency appears to be beneficial for the trust calibration process (Yang et al., 2017). The transparency that dissimilarity scores add to an AI system is unknown. Exploring whether participants find this information helpful in their decision-making can inform more about trust calibration.

To better predict the role of individual differences and to account for the uniqueness of each face stimulus, the current study will make use of mixed effects modelling. By treating participants and face stimuli as random variables, results can be generalised to a larger population of human operators of technology and other face materials. Mixed effect modelling may be particularly useful in studies that make use of repeated measures (Baayen et al., 2008). The current study adds to the literature by taking into account variations that occur within trials and participants.

Experiment 2 focused only on using AI with limited reliability as Experiment 1 failed to find differences in trust ratings between groups using AI with high or low reliability. Previous research has found that providing participants with inconsistent labels reduces accuracy on a face-matching task (Fysh & Bindemann, 2018a). Experiment 2 examined trust in AI by analysing trust ratings on a given label, which could be consistent or inconsistent with the trial. To reflect human-AI interactions in real-life situations, the current study made use of the long version of Kent Face Matching Test (KFMT), with infrequent identity mismatches (Fysh & Bindemann, 2018c) and examined face-matching performance on consistently and inconsistently labelled trials, focusing primarily on ‘false’ alarms of the

system. By treating participants and face stimuli as random variables, results can be generalised to a larger population of human operators of technology and other face materials. Mixed effect modelling is particularly useful in studies that make use of repeated measures designs (Baayen et al., 2008). The current study adds to the literature by taking into account variation that occurs across trials (stimuli) and participants.

Whether the lack of significant findings in percentage accuracy and sensitivity in Experiment 1 was related to individual differences remains unanswered. Experiment 2 will address this limitation by using a within-participant experimental design and taking into account participants' propensity to trust. Given that Experiment 1 did not find significant findings in trust ratings on the system, Experiment 2 will continue to use a similar set-up but analyse trust in AI on a trial-by-trial basis, as opposed to the system on the whole.

In the current experiment, performance was examined by comparing the accuracy of consistently and inconsistently labelled trials in conditions where AI support is provided with or without explanation. The goal of including both consistently and inconsistently labelled trials was to examine the impact of using imperfect AI. As performance may be related to trust, subjective trust ratings are expected to also be higher in consistently than inconsistently labelled trials, particularly in conditions where an explanation is provided. Reflecting the lower frequency of mismatches at the border, the current study mirrored a real-life application of face-matching by using an uneven number of match and mismatch trials to examine the usefulness of algorithm support. Participants' propensity to trust automation was assessed to explore the possible influence of this personality trait on trust calibration.

4.2.2 Method

Participants.

Participants were recruited through the online platform *Prolific* (*www.prolific.com*). There was a total of 36 participants (mean age: 25.44) who completed the experiment. There were 32 females and 4 males, and all self-reported to be White/Caucasian as their ethnicity. Participants were compensated using *Prolific's* payment system, which was £7.00 per hour.

Materials.

KFMT: Face images were taken from the Kent Face Matching Test (Fysh & Bindemann, 2018). A total of 108 image pairs were used with 100 matches and 8 mismatches in the experiment, split equally in each block of trials. Four of the image pairs were used in the practice trials. The side at which the student ID image and portrait photo appeared were randomised. The order in which images were presented was also randomised.

Face-recognition: The same face-recognition algorithm as in Experiment 1 was used. All faces from the KFMT dataset were processed using the library and a list of dissimilarity scores was obtained for each image pair. Four pairs of images were used in the practice trials. There were 54 images from the KFMT in each block of trials and 25 match trials were inconsistently labelled as 'different'. Images with the highest dissimilarity scores were inconsistently labelled to be different.

Propensity to Trust Automation: These questions aimed to assess propensity to trust automation (Merritt, 2011) and were adapted to be more context-specific for Automated Facial Recognition. Questions required a response between strongly disagree and strongly

agree, scored from one to six. The survey consisted of 6 items and the value for Cronbach's Alpha for the survey was $\alpha = .87$.

PsychoPy interface: The experiment was designed on PsychoPy (Peirce et al., 2019) and was similar to the layout used in Experiment 1, with the addition of dissimilarity scores in the condition where additional AI information was provided.

Design.

The study made use of a within-participant design and each experimental condition was given to participants in a counterbalanced order. The independent variables were the availability of dissimilarity scores (with or without dissimilarity scores) and consistency of labels (consistent or inconsistent). The dependent variables were trust and face-matching decisions. Trust was measured using subjective ratings made on a scale of 0-100% with anchor points at 20%, 40%, 60% and 80%.

Procedure.

Participants were asked to read the Participant Information Sheet before proceeding to the online experiment on *Pavlovia* (*https://pavlovia.org/*) and were required to read and respond to a series of statements regarding consent before beginning by pressing the relevant keyboard responses.

Participants were given the context and explanation of the function and purpose of automatic facial recognition and asked to answer a series of questions that assessed their propensity to trust automation. Participants were informed that they would be presented with two face images in each trial and that their task was to compare the two images and decide whether they belonged to the same person or different people. Instructions included the idea that an AI made decisions based on a dissimilarity score and a given threshold. Participants were then given practice trials which provided the participant with two

examples of a match and a mismatch trial. In each block of trials, there were two famous face pairs, included as attention checks. These were iconic pictures of politicians that participants were expected to recognise.

Participants were required to match faces with AI advice in conditions with dissimilarity scores. All participants were given AI support in the form of 'same' or 'different'. Additional information in the form of a dissimilarity score was presented at the same time as the AI advice in one block of the trials. Participants were informed that the advice was not guaranteed to be correct. A slider also appeared prompting responses on trust ratings, ranging from 0-100% on a continuum, with 0 to the left of the scale and 100 to the right with anchor points at 20%, 40%, 60% and 80%. Following a response on the trust scale, a confirm button appeared allowing the participant to finish the current trial and proceed to the next. Debriefing questions were included at the end and aimed to explore further factors related to the experience of being a participant.

4.2.3 Results

Of the 36 participants, four participants did not respond correctly on all four famous face trials but none responded incorrectly on more than one therefore all participants were included for data analysis. Famous face trials were only included to verify the engagement of participants in the experiment and were not included for further analysis. Non-match trials were included to mimic real-life situations but were excluded from the analysis.

Performance.

The percentage accuracy in each block of trials was obtained by the number of correct trials divided by the total number of trials in a given type of condition. Table 3 is the summary of results and contains the mean accuracy for consistently and inconsistently labelled trials in each condition, along with their standard deviations.

Table 4

Mean Percentage Accuracy (SD in brackets) with and without Dissimilarity Scores

	Condition	
Label	Dissimilarity Scores	No Dissimilarity Scores
Consistent	84.77 (11.78)	85.11 (15.38)
Inconsistent	55.22 (23.30)	57.00 (23.58)

To examine Condition and Label as predictors of the face-matching decision, a *binary logistic* mixed effects regression model was built in R using the *lme4* package (Bates, Mächler, Bolker, Walker, 2015). Condition and Label were entered as fixed factors (using mean-centred deviation coding) and Participants and Items were included as random factors. The model was specified as:

*Decision ~ Condition * Label +*

*(1 + Condition * Label | Participant) +*

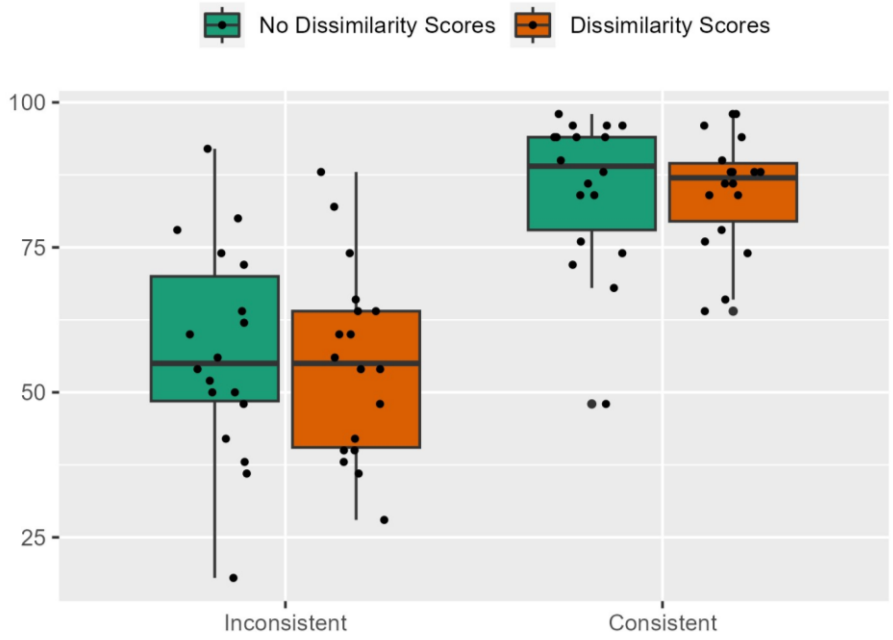
*(1 + Condition * Label | Items)*

The model employs the maximal random effects structure justified by the design, appropriately taking into account that Condition and Label were manipulated both within participants and within items (see Barr, Levy, Scheepers, & Tily, 2013).

There was a significant main effect of Label (*Est.* = -2.14, *SE* = 0.297, *Wald Chi-Square* (1) = *51.88*, $p < .001$) indicating that accuracy reliably decreased with inconsistent rather than consistent labels (See Table 3). The main effect of Condition ($p = 0.399$) and the Condition X Label interaction were not significant ($p = 0.649$). Figure 12 is an illustration of the percentage accuracy in each condition.

Figure 12

Boxplots of percentage accuracy in each condition



Trust.

Table 5 is a summary of the results on trust ratings.

Table 5

Mean Trust Ratings with and without additional AI explanation

	Condition	
Label	Dissimilarity Scores	No Dissimilarity Scores
Consistent	67.39 (23.29)	70.08 (23.87)
Inconsistent	45.83 (25.66)	45.62 (27.50)

Condition and Label were entered as mean-centred predictors of Trust Ratings, also with participants and items as random factors:

*Trust ~ Condition * Label +*

*(1 + Condition * Label | Participant) +*

*(1 + Condition * Label | Items)*

Trust Ratings were analysed as a continuous variable and a standard linear mixed effects regression approach was used. Results indicated no significant interaction (*Est.* = 2.899, *SE* = 3.616, *Wald Chi-Square* (1) = 0.643, *p* = .423) and there were no main effects of the Condition (*Est.* = -1.243, *SE* = 1.781, *Wald Chi-Square* (1) = 0.487, *p* = .485). Trust ratings were generally higher in consistently labelled trials than in inconsistently labelled trials (*Est.* =23.012, *SE* = 3.286, *Wald Chi-Square* (1) = *49.047*, *p* < .001).

Propensity to Trust.

The mean propensity to trust rating was calculated for each participant and to examine whether propensity to trust and subjective trust were predictive of decision, a model was built by adding the two measures of trust and consistency as fixed factors, with participants and items as random effects as specified below:

*Decision ~ Label * Trust * Propensity +*

*(1 + Label * Trust | Participant) +*

*(1 + Label * Trust * Propensity | Items)*

There was a main effect of Label (*Est.* = -3.788, *SE* = 0.504, *Wald Chi-Square* (1) = *56.409*, *p* < .001), and trust (*Est.* = 1.198, *SE* = 0.233, *Wald Chi-Square* (1) = *26.506*, *p* < .001). However, Propensity to Trust alone was not predictive of decision (*Est.* = 0.209, *SE* = 0.230, *Wald Chi-Square* (1) = *0.828*, *p* = .363). Trust X Label interaction was also significant (*Est.* = -8.817, *SE* = 0.988, *Wald Chi-Square* (1) = 79.684, *p* < .001).

4.2.4 Discussion

Experiment 2 aimed to gain a better understanding of how trust is involved in the interaction between human face-matching decision-makers and facial recognition systems by examining trust in AI's consistent or inconsistent advice in each trial. Previous research exploring the interaction between humans and computers in an unfamiliar face-matching task has confirmed that people are biased by inconsistent labels where match identities were labelled as different and mismatched identities as the same (Fysh & Bindemann, 2018). Experiment 2 aimed to contribute by examining variations in trust ratings on a trial-by-trial basis and taking into account the role of participants' propensity to trust.

Understanding the influence of trust in AI can foster appropriate levels of trust (Hoff & Bashir, 2015), hence facilitating the interaction between humans and AI. With the aim to explore trust calibration to improve performance, Experiment 2 investigated whether presenting dissimilarity scores with AI labels had an influence on decision-making and trust in a face-matching task. The hypothesis was that face-matching decisions and trust can be predicted by the consistency of labels and the availability of AI scores.

In line with expectations, the consistency of the label provided by AI was a significant predictor of face-matching decisions. Higher percentage accuracy for consistently labelled trials than inconsistently labelled trials also suggests that participants were biased by the labels as they followed incorrect advice given by the AI. However, contrary to the hypothesis, results showed that AI support accompanied by AI dissimilarity scores was not predictive of face-matching decisions. As suggested by insignificant differences in percentage accuracy between the two conditions, the present findings indicate that simply presenting AI dissimilarity scores was insufficient in improving face-matching decisions when given an AI that provided inconsistent labels. Participants were only provided with brief instructions on how to interpret the dissimilarity scores and this may not have been sufficient.

Propensity to trust automation is the tendency to be trusting of machines in general (Merritt & Ilgen, 2008) and was assessed as a separate measure of trust. Results indicated that propensity to trust, measured on a continuum, was not a significant predictor of face-matching decisions. On the other hand, trust ratings were predictive of responding ‘match’ or ‘non-match’. There was also an interaction between subjective trust and the consistency

of labels. This suggests that AI performance-related trust was more influential in behaviour than personal characteristics.

Trust ratings were predictive of face-matching decisions in the form of 'match' or 'non-match', confirming that trust is involved in the interaction between the AI and human decision-makers. This also suggested that trust that is learned throughout the interaction with an AI system, is more predictive of behaviour than dispositional factors. However, the current experiment is unable to provide direct evidence of actual learning having taken place. Trust also interacted with the consistency of the label to predict face-matching decisions which indicated that participants were able to distinguish between the two types of trials to help inform behaviour.

The lack of a significant effect of propensity to trust can be explained by the difficulty of measuring propensity to trust. Previous studies have used different measures of propensity to trust which vary in reliability and predictive validity of trustworthiness of automation and actual trusting behaviour (Jessup et al., 2019). The questions used in the current experiment were specifically adapted to be relevant to facial recognition algorithms and were shown to demonstrate relatively high internal consistency (Tabererg, 2017), but these questions have not been previously validated.

Experiment 2 did not find that presenting AI scores with advice was predictive of face-matching decisions. It could indicate that the transparency of AI simply does not aid face-matching decisions as it does with other tasks in other domains. The current experiment presented AI scores which indicated the AI-estimated dissimilarity between images, however, it is unclear how this information was used or understood by participants. Future studies could examine the features of transparency specific to face recognition that

could be useful, for instance, by improving both interpretability and uncertainty awareness of AI (Tomsett et al., 2020). Transparency can take the form of explanations, uncertainty estimates or performance metrics (Zerilli et al., 2022). Trust calibration designs should be specific to the task.

Alternatively, it can be argued that displaying AI scores did not improve AI transparency due to a lack of understanding or interest in the task, unlike human decision-makers in applied settings. The way AI information is used may also be related to initial familiarity with the task and AI (Schaffer et al., 2019). Studies can further examine different ways to present AI scores to better communicate the system's capabilities. The results of the current experiment on trust replicate the finding that trust towards classifiers appears to be based primarily on the system's performance (Ingram et al., 2021).

In summary, the results of Experiment 2 suggested no obvious benefit of introducing explanatory information in the form of dissimilarity scores for users of AI. The performance of the AI appeared to be the primary predictor of face-matching decisions. Participants tended to conform to the decisions made by AI, despite being informed that some advice may not be accurate. Trust calibration was proposed to be able to solve this issue. By providing additional information such as dissimilarity scores, participants were expected to trust AI advice when it is reliable and transparent and to use alternative resources, such as their own expertise, when the system is not reliable.

Experiment 2 investigated the possibility of trust calibration and verified the role of trust in human-AI interaction. When trust was analysed on a trial-by-trial basis, the second experiment found that trust differed significantly between consistently and inconsistently labelled trials. Results also showed that the consistency of the label was predictive of face-

matching decisions. However, providing additional AI information in the form of dissimilarity scores did not improve performance or influence trust.

To conclude, Experiment 2 showed that displaying dissimilarity scores did not impact decision-making or trust. The consistency of AI labels appeared to be the main predictor of both face-matching decisions and trust. More research is required to examine how the transparency of automated facial recognition systems can be enhanced. These results confirm findings that inconsistent labels bias decision-making, in line with existing research on human-AI interactions in face matching. The findings of the current study provide support for the involvement of trust in human-AI interactions in the context of face-matching. Future studies could continue to explore the calibration of trust as a way to facilitate the interaction between humans and AI in face-matching tasks.

Chapter 5: Expertise in Face-matching and Trust in AI

5.1 Experiment 3

5.1.1 Introduction

Face matching is a cognitive process that involves comparing two face images simultaneously and determining whether they belong to the same identity. Research has consistently demonstrated that humans have a poor ability to match unfamiliar faces, compared to familiar faces (Megreya & Burton, 2006). For example, studies have revealed that individuals often make errors when attempting to identify unfamiliar faces captured in CCTV images. However, their performance significantly improves when they are familiar with the individual (Bruce et al., 2001; Burton et al., 1999). Moreover, when asked to sort face images into different identity piles, people tend to overestimate the number of unique identities present, but their accuracy improves when they are familiar with the faces in the photographs (Jenkins et al., 2011). The limitations in unfamiliar face matching are particularly problematic in applied settings that require verifying identities, such as in forensics settings and border control, as inaccuracies under these circumstances have severe consequences. By developing strategies and techniques to improve face matching, face-matching research aims to enhance the accuracy and reliability of identity verification processes by reducing the risk factors associated with incorrect identifications in applied scenarios.

Professionals who perform unfamiliar face-matching as part of their daily job have also shown variable ability in face-matching. Studies have revealed insights into the relationship between expertise and face-matching performance, showing that forensic examiners exhibited a slight advantage over untrained students, particularly at longer

exposure durations (White, Phillips, Hahn, Hill & O'Toole, 2015). Similarly, passport officers outperformed students in matching photos to official IDs but also took significantly more time to match unfamiliar faces. Interestingly, the length of time employed as an officer did not predict accuracy, suggesting that experience might not be the sole factor determining performance (White, Kemp, Jenkins, Matheson & Burton, 2014). These studies highlight the complex relationship between expertise and face-matching performance. Previous research has suggested that professional experience does not necessarily lead to expertise as facial reviewers who self-reported at least one year of professional ID card screening experience were just as susceptible to the low prevalence effect in face matching compared to non-professionals (Weatherford et al., 2021). In low prevalence conditions, non-matching faces were more likely to remain undetected among both groups.

Another study has found that untrained students had more false positives than false negatives compared to forensic experts, as experts tended to be more careful with their conclusions when image quality was low (Norell, Lathem, Bergstrom, Rice, Natu & O'Toole, 2015). Despite mixed findings on the accuracy of face-matching, one consistent and robust finding across these studies is that trained facial comparison experts and novices indeed approach face-matching tasks differently. The specific strategies and decision-making processes employed by experts compared to novices can significantly impact their performance and outcomes in face-matching tasks.

Research has also explored whether facial recognition technology alone or in combination with humans could achieve higher accuracy than humans alone. By combining the collective capabilities of both humans and AI algorithms, facial recognition systems can be optimized for enhanced performance. Research has also considered factors like

experience and individual differences. *Super recognizers* are individuals with extraordinarily high levels of accuracy in both face perception and recognition (Russell et al., 2009). A recent study comparing the performance of professionals and super recognisers with undergraduate students and algorithms has revealed that fusing the highest-performing group with the best-performing algorithms yielded the greatest level of accuracy (Phillips et al., 2018). This finding highlights the potential of human-AI teams to significantly improve face-matching accuracy. At present, many verification systems, such as those used in border control, rely on human operator oversight to make face-matching decisions based on algorithm outputs.

Research examining the effects of presenting algorithm results in the form of labels such as “same”, “different” and “unresolved” has found that prior decisions made by algorithms can influence the subsequent face-matching decisions made by human operators. When face pairs were inconsistently labelled as ‘same’ or ‘different’, accuracy decreased by diverting attention away from the face images (Fysh & Bindemann, 2018). Mimicking the higher frequency of matched to mismatched cases at the border, this study illustrated the potential errors in the human-AI interaction at passport control and suggests that inaccurate text cues may be detrimental to performance. Likewise, another study has shown a similar bias as participants were mostly correct when no prior information was given but introducing labels biased their certainty judgements, with no effects on their ability to discriminate between a match and a mismatch, as observers with high decision threshold were more likely to classify face pairs as different and observers with low thresholds were more likely to respond same (Howard, Rabbitt & Sirotin, 2020).

Issues also arise in other forms of face-matching verification systems. Research has shown that in facial recognition systems where human decision-makers are presented with a list of candidate images ranked by their similarity to a target image, the number of candidate matches presented to reviewers significantly affected performance in a face-matching task for unfamiliar faces. The study found that longer lists with 100 images produced more false alarms, lower confidence ratings and increased response latencies in both experienced and inexperienced facial reviewers (Heyer, Semmler & Hendrickson, 2018). To address these issues, automated systems should be designed in a way that minimises bias and supports the decision-making process in a meaningful way. Optimising the human-AI interaction has the potential to facilitate a more reliable and efficient face-matching process.

Trust in automation and decision support systems

Improving the collaboration between humans and AI could involve calibrating the levels of trust and interaction based on the actual reliabilities of the AI system. Trust is a psychological concept that is involved in all human-AI interactions and appears to be an area that is overlooked in the human-AI face-matching literature. Trust can be defined as an attitude or expectation of a favourable response (Rotter, 1967). Another commonly used definition is that trust is an intention and involves an element of vulnerability (Mayer et al., 1995). Lee and See (2004), in an influential integrated review of early research on trust and reliance on automation, conceptualised trust as an attitude leading to the intention of specific behaviour, suggesting that trust can be measured as a behavioural outcome. Incorporating trust as a factor in human-AI interaction could enhance the overall performance of face-matching tasks. By considering the actual reliabilities of an AI system,

human operators can make informed decisions on when and how to make use of the AI's outputs. Examining trust in the human-AI collaboration has the potential to aid the interaction and produce positive benefits.

Propensity to trust can be described as a personality trait that represents the general tendency of a person to trust another (Mayer et al., 1995b). Adapted to trust in automation, researchers have devised several measures to predict perceived trustworthiness and behavioural trust (Jessup et al., 2019). In general, people are often willing to trust novel technologies (Dzindolet et al., 2003). Initial trust in automation is based primarily on faith and rapidly decreases after seeing system errors as dependability and predictability become key contributors to trust (Madhavan & Wiegmann, 2007). This highlights the importance of building and maintaining trust in AI systems.

The current study explored the role of trust in a face-matching context to examine potential biases in decision-making when an AI makes errors. Specifically, we focused on uncovering differences in trust behaviour between two groups: professionals and novices in face-matching. Professionals included facial reviewers, facial examiners and police investigators, with varying levels of expertise. Expertise in the current study refers to superior face-matching abilities that are acquired through on-the-job training and experience. Facial reviewers are professionals who make quick face-matching decisions at large volumes while facial examiners tend to make more detailed forensic comparisons of facial images at longer durations (White et al., 2015). Alternatively, participants had the option to self-identify as a police investigator. The study excluded super-recognisers, as how expertise is acquired for this group of high achievers appears to be less clear. The current research will examine the effect of professional expertise on trust and performance by

comparing differences between professionals and novices in face-matching behaviour when given AI support. For the current study, novices will be participants who do not fall into any of the following categories: facial reviewers, facial examiners, police investigators and super recognisers.

Drawing on research from other domains, it can be seen that subject matter expertise can alter trust in automation (Hoff & Bashir, 2015), as professionals and novices interact with AI and automated systems differently. In a perceptual task which required participants to detect whether a weapon was present in a series of X-ray images of cabin baggage, Chavaillaz and colleagues (2019) found an increase in performance for novices but not for professionals when assisted by a diagnostic aid. Another study has shown that participants with experience and understanding of a specific domain were more reluctant to rely on automation, such as in operating agricultural vehicles (Sanchez et al., 2014). Not calibrating trust appropriately could result in the disuse and misuse of automation, referring to the over-reliance and under-reliance of automation (Parasuraman & Riley, 1997). Trust calibration involves aligning a user's trust with the automation's true capabilities (Lee & See, 2004). Understanding trust behaviour between professionals and novices would assist trust calibration and may help to enhance face-matching performance. By investigating the impact of AI assistance on face-matching and exploring the influence of face-matching expertise, the aim of the current study is to examine whether distinct needs and behaviours of professionals and novices can be accommodated to optimise the design and usage of AI systems.

The focus of the current study was on the interaction between human face-matching decision-makers and facial recognition systems, examining the roles of professional face-

matching expertise, propensity to trust, and actual trust in the AI recommendation as well as confidence. Expertise and propensity to trust are person-specific variables, whereas trust in the AI recommendation and confidence were established on a by-trial basis. The study will recruit face-matching professionals and age and gender-matched novices without specific expertise in face-matching. Professionals are expected to have acquired their expertise in face matching from professional training and experience and will be asked to self-report the number of years in their current occupation.

We expect professionals to be less reliant on AI support, which should be reflected in lower "trust in AI" ratings. We also expect there to be a difference between professionals and novices in terms of confidence ratings, though the exact nature and extent of the influence are less clear. For instance, it is unclear whether expertise could lead to overconfidence by overestimating their judgments as correct or believe that they outperform their peer (Sanchez & Dunning, 2023), or reduced confidence due to realistic assessments of the difficulty of the task. Individuals often have limited awareness of their face recognition abilities, likely due to a lack of training and feedback outside controlled laboratory settings (Bindemann et al., 2014). In particular, naïve participants are less accurate in their judgements of face-matching performance compared to participants who have been previously informed of their abilities (Bobak et al., 2019). Exploring the relationship between expertise, trust and face-matching performance is expected to provide insight into the decision-making process and have significant implications for optimising human-AI interactions.

5.1.2 Method

Participants.

The study recruited professionals from a face-matching related professional association, specifically targeting facial reviewers, facial examiners, and police investigators. Only professionals who fell into one of these categories were asked to proceed to the experiment, where they were asked to self-report their specific category. Other information collected were age, gender ethnicity, employment status and the number of years of experience in the occupation. Non-expert (novice) participants were recruited through Prolific (*https://www.prolific.com)* and were expected to lack experience in face-matching. Novice participants were asked to only proceed with the experiment if they did not fall into any of the above categories and were selected based on their age and gender to ensure matched control for comparison purposes between the groups.

Materials.

GMFT2: The test consisted of image pairs that required participants to match identities across variations in head angle, pose, expression and subject-to-camera distance (White et al., 2021).

AI Labels: Images were accompanied by labels of ‘same’ or ‘different’. Participants were told that these were outputs of an AI system and that the reliability of the system is unknown. Twenty images out of forty were inconsistently labelled.

Propensity to trust questions: These items were modified versions of the Propensity to Trust Technology scale, designed to assess individuals' attitudes toward technology adoption (Jessup et al., 2019). The questions were intended to assess participants' likelihood of utilizing the AI. An example item was "I think it’s a good idea to rely on AI for help."

Participants were required to select a response on a continuous scale ranging from strongly disagree to strongly agree.

Design.

The experiment employed a mixed design with participant group (professionals, novice) as between-participant/within-item and AI labels (consistent, inconsistent) as within-subject/within-item factor.

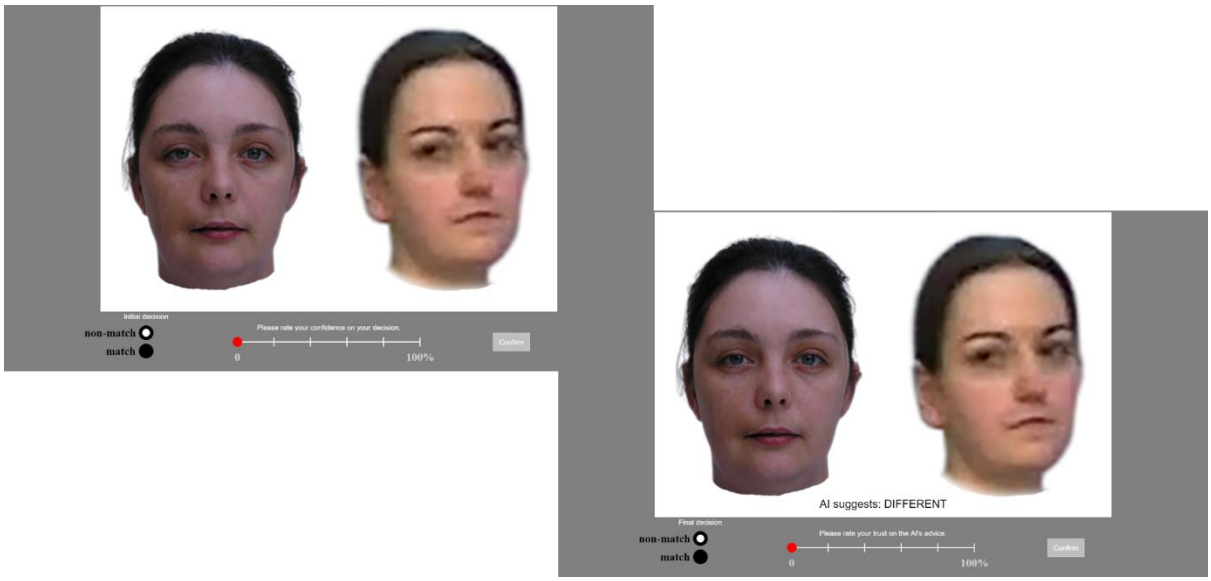
Procedure.

Participants were asked to complete the GFMT2, within a single experimental block containing 40 trials consisting of an image pair of the same or different identity. On each trial, they were given an image pair and asked to decide whether they were the same person or different people by responding 'match' or 'non-match' and to rate their confidence in the decision that they made using a continuous scale ranging from 0-100. They were then given AI advice and asked to provide a trust rating on the AI support, also on a continuous slider from 0 to 100. Participants were given the opportunity to decide again whether the faces belonged to the same person or different people by responding 'match' or 'non-match'. Participants had to respond fully and click 'confirm' before proceeding to the next trial. Figure 1 illustrates the trial sequence: making a decision without AI advice, making a decision with AI advice and then the next trial.

Figure 1

Layout and sequence of a trial: Participants first provide a match/mismatch decision and a corresponding confidence rating. Next, they see a "same/different" AI recommendation and

have to provide a trust rating on it.



Participants

There was a total of 56 participants: 28 professionals and 28 novices. Table 1 is a summary of the demographic information of participants.

Table 1

Sociodemographic Characteristics of Participants

Characteristic	Professionals		Novices		Full sample	
	n	%	*n*	%	*n*	%
Gender						
Female	14	50	14	50	28	50
Male	14	50	14	50	28	50
Occupation						
Facial Examiner	17	61	0	0	17	30

Facial Reviewer	3	11	0	0	3	5
Police Investigators	2	7	0	0	2	4
Other	6	21	28	100	34	61
Ethnicity						
White/Caucasian	27	96	28	100	55	98
Other	1	4	0	0	1	2
Employment						
Full-time Employment	24	86	16	57	40	72
Part-time Employment	2	7	3	11	5	9
Not in Employment	0	0	1	4	1	2
Homemaker	0	0	4	14	4	7
Self-employed	1	4	2	7	3	5
Prefer not to say/Other	1	4	2	7	3	5

Note. Participants were on average 40.29 years old (*SD* = 8.96), and participant age did not differ by Group as “novices” were recruited as age- and gender-matched controls. The mean number of years in the current occupation was 7.43 for professionals and 10.38 for novices.

5.1.3 Results

Descriptive statistics showed that professionals exhibited a higher percentage accuracy, with a mean accuracy of 95.36% on consistently labelled trials and 71.79% on inconsistently labelled trials. In comparison, novices achieved a mean accuracy of 87.50% on consistently labelled trials and 68.21% on inconsistently labelled trials.

Measures of performance were sensitivity and bias, *d*’ and *c*, calculated with the R package "psycho" (Makowski, 2018). Table 6 is a summary of the results indicating the

higher sensitivity of professionals on both consistent and inconsistent trials and similar bias in both groups of participants. Figures 13 and 14 show the distribution of *d*' and c for each condition.

Table 6

Mean and standard deviations of d' and c on each type of trial for each group

	d'		***c***	
Group	M (SD)		M (SD)	
	Consistent	Inconsistent	Consistent	Inconsistent
Professionals	2.93 (0.72)	1.21 (1.09)	-0.12 (0.35)	0.00 (0.46)
Novices	2.23 (0.73)	0.99 (0.73)	-0.01 (0.35)	0.01 (0.40)

Figure 13

Average d' for professionals and novices

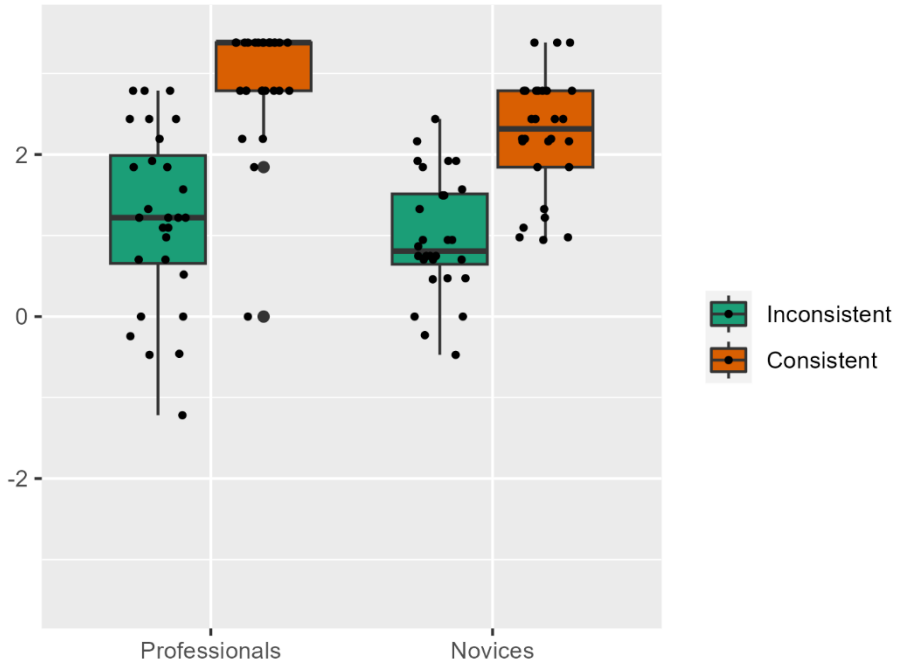
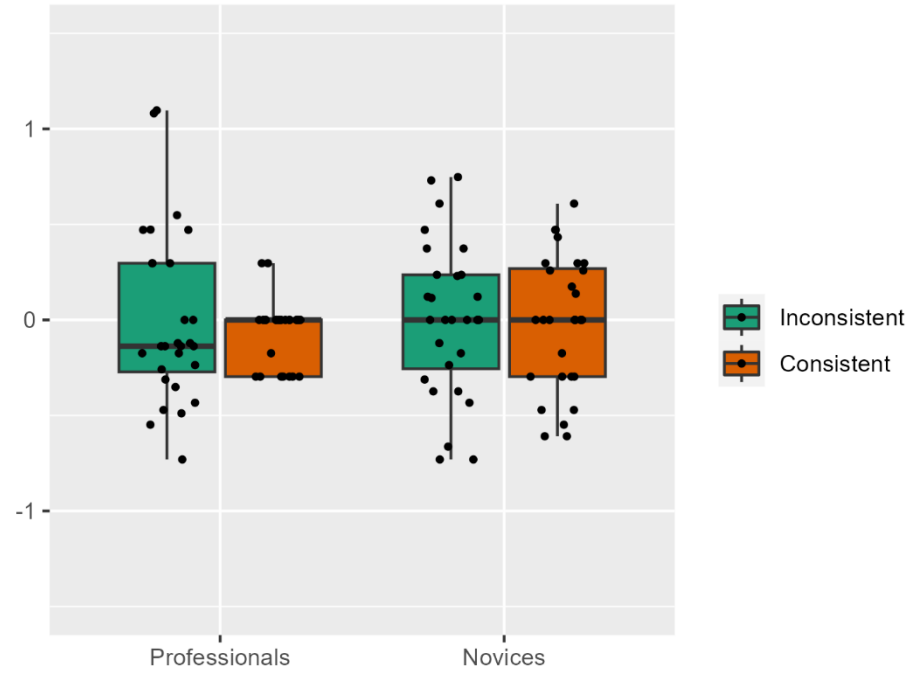


Figure 14

Average c for professionals and novices



To examine the effects of Group (professionals or novices) and Label (consistent or inconsistent) as predictors of decision outcome, a mixed effects model was built in R using the *lme4* package (Bates, Mächler, Bolker & Walker, 2015). Group and Label were entered as fixed factors using mean-centred deviation coding and Participants and Items were included as random factors. Taking into account that only consistency of labels was manipulated within participants, whereas both Group and Label varied within items, we used the following model structure (with "Outcome" being a binary DV taking the values "correct" = 1 or "incorrect" = 0 on any given trial):

*Outcome ~ Group * Label + (1 + Label | Participant) + (1 + Group * Label | Items)*

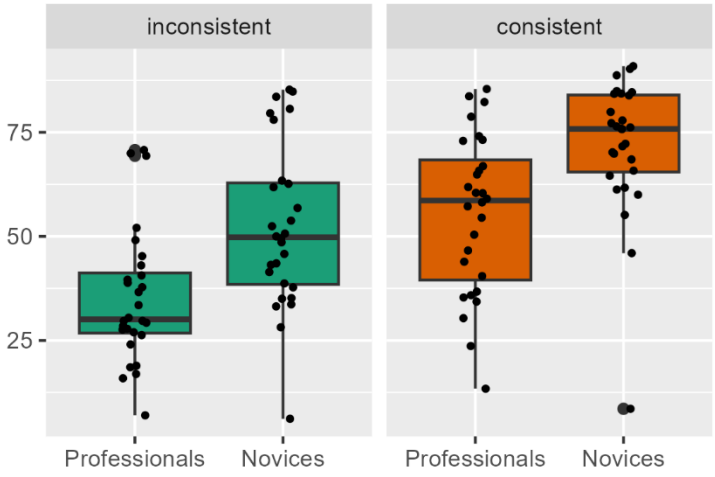
There was a significant main effect of both Group (*Est.* = -0.603, *SE* = 0.180, *Wald Chi-Square* (1) = *11.209 p* < .001) and Label (*Est.* = -1.901, *SE* = 0.316, *Wald Chi-Square* (1) = *36.250*, *p* < .001). Their interaction was also significant (*Est.* = 0.770, *SE* = 0.360, *Wald Chi-Square* (1) = *4.559*, *p* = .033), indicating that the decision outcome can be predicted by the consistency of the label and professional status of participants. Further analyses indicated a difference between professionals and novices when the trials were consistently labelled (Est. = -0.988, SE = 0.318, *p* = .002) but not in inconsistently labelled trials (Est. = -0.219, SE = 0.169, p = .196). The influence of label consistency was evident in both professionals (Est. = -2.285, SE = 0.375, *p* <.001) and novices (Est. = -1.516, SE = 0.352, *p* <.001).

Trust

Trust ratings were collected for each trial and averaged by participants for comparisons between groups. Experts generally had lower trust ratings than novices in both conditions. On consistently labelled trials, both professionals (M = 55.38, SD = 27.15) and novices (M = 71.67, SD = 24.99) showed higher trust ratings compared to inconsistently labelled trials, where professionals (M = 35.14, SD = 25.19) and novices (M = 52.26, SD = 29.32) had lower trust ratings. Figure 15 displays the differences in trust ratings between the consistently and inconsistently labelled trials.

Figure 15

Differences in mean trust ratings between professionals and novices



The mean propensity to trust rating was 3.94 for professionals and 3.49 for novices. To examine the propensity to trust as a predictor of trust ratings, analysis was carried out with the propensity to trust for each participant included as a fixed factor to examine its influence on trust ratings, alongside the variables Group and Label; Trust itself was treated as a continuous DV (using a standard linear mixed effects model):

*Trust ~ Group * Label * Propensity + (1 + Label | Participant) + (1 + Group * Label * Propensity | Items)*

Results showed a significant effect of Label (*Est.* = -20.38, *SE* = 2.37, *Wald Chi-Square* (1) = *73.55, p* < 0.001), suggesting that inconsistently labelled trials tended to result in lower trust ratings compared to consistently labelled trials. The main effect of Group was significant (*Est.* = 18.63, *SE* = 1.42, *Wald Chi-Square* (1) = *171.04*, *p* < .001).

The main effect of Propensity to trust was significant (*Est.* = 3.326, *SE* = 0.804, *Wald Chi-Square* (1) = *17.12, p* < .001) but further modulated by a Group X Propensity interaction (*Est.* = 4.89, *SE* = 1.41, *Wald Chi-Square* (1) = 11.98, *p* < .001), indicating a different impact of propensity to trust on trust ratings between professionals and novices. Further analyses showed that the effect of trust propensity on trust was significant for novices (*Est.* = 7.485, *SE* = 1.252, *Wald Chi-Square* (1) = 35.76, *p* < .001), but not for professionals (*Est.* = 1.141, *SE* = 1.515, *Wald Chi-Square* (1) = 0.569, *p* = .450).

Decision Changes

Incorrect changes of decision were higher for novices (9.43% on mismatch trials and 15.45% on match trials) than professionals (8.80% on mismatch trials and 14.77% on match trials). The following binary logistic mixed effects model was used to examine potential differences in changes of decision (1 = “change”, 0 = “no change”) after AI advice was presented:

*Change ~ Group * Label * Confidence + (1 + Label * Confidence | Participant) + (1 + Group * Label * Confidence | Items)*

There was no main effect of Label (*Est.* = 0.242, *SE* = 0.356, *Wald Chi-Square* (1) = *0.462*, *p* = .496), but a main effect of Group (*Est.* = 0.669, *SE* = 0.237, *Wald Chi-Square* (1) = *8.128*, *p* = .005) and Confidence (*Est.* = -0.763, *SE* = 0.184, *Wald Chi-Square* (1) = *17.121*, *p* < 0.001). There were also no interactions between any of the variables. Changes of response from a correct decision to an incorrect decision on inconsistently labelled trials were more frequent in novices than professionals (See Appendix for Sankey diagrams displaying the flow of decisions for professionals and novices).

5.1.4 Discussion

The present study investigated the role of professional expertise in unfamiliar face-matching when given AI assistance. Participants were categorised into professionals versus novices based on occupational background and were asked to match unfamiliar faces. They were presented with AI recommendations that were either correct (consistent) or incorrect (inconsistent). There were several hypotheses relating to the role of expertise in face-matching. These hypotheses focused on whether the enhanced performance of professionals persisted when provided with unreliable AI support, whether there were differences in trust ratings between the groups and if these could be attributed to an individual propensity to trust measured using an adapted propensity to trust technology scale. Confidence ratings were also collected to explore whether reliance on AI could be explained in terms of levels of confidence for any given trial along with trust in the AI.

It was expected that professionals, due to their professional training and experience, would be more accurate in their face-matching decisions compared to novices. Previous research has suggested professionals to be more conservative in their responses compared to novices (Norell, Lathem, Bergstrom, Rice, Natu & O'Toole, 2015), particularly when image quality is low. This study explored whether this finding persists when face-matching with AI of limited reliability. Sensitivity and bias were used as measures of performance to gain insights into these aspects of unfamiliar face matching.

The findings on sensitivity highlight a difference between professionals and novices in their ability to discriminate a match from a mismatched identity face pair, particularly in identifying true identity matches. This is consistent with the hypothesis that expertise plays a crucial role in face-matching performance. Previous research has suggested that certain

groups of professionals, such as facial examiners, perform better than novices without training in facial image comparison (White et al., 2015). Our findings suggest that the superior performance of professionals persists even in the presence of potentially unreliable AI assistance that provides both consistent and inconsistent outputs. However, inconsistent labels reduced the sensitivity of both novices and professionals, perhaps suggesting that both professionals and novices conform to incorrect AI advice. Results using mixed-effects modelling further showed that expertise was not predictive of correct face-matching decisions on inconsistently labelled trials, showing that incorrect AI advice affects both groups.

Our findings on decision changes showed that professionals were less likely to change their decision following AI advice compared to novices. This could be attributed to higher self-confidence in professionals. However, there were no significant interactions, suggesting that the effects of confidence and group were more likely to be independent. Research on automation bias in healthcare has found that decision switches were higher in those with less clinical experience (Goddard et al., 2014). While the current study was on face-matching, this similarity demonstrates the broader relevance and applicability of the findings in other human-AI interactions.

Furthermore, there was no significant difference in bias indicating it does not appear to be affected by AI and incorrect AI advice. Levels of bias were similar in both groups suggesting that the role of AI assistance might have moderated the influence of expertise on decision tendencies as professionals were not more conservative in their responses as expected. Results indicate no changes in bias after seeing inconsistent labels in both

professionals and novices. This could be related to the lack of consequence in making an error in the current experiment, in contrast to real-world scenarios.

Additionally, it was anticipated that professionals would exhibit greater self-reliance in face-matching tasks, leading to reduced dependence on AI assistance while novices were expected to rely more on AI support due to their lack of experience in face-matching. Trust ratings were collected for the provided AI labels, indicative of potential differences between professionals and novices in their reliance on AI. Findings show that trust ratings were generally higher for novices than professionals, and both professionals and novices were able to distinguish between consistently and inconsistently labelled trials, as trust ratings were generally higher for consistently than inconsistently labelled trials. However, the lack of interaction suggests that the impact of label consistency was similar for both groups. These results suggest that novices were more trusting of AI and incorrect AI advice affects both groups to a similar extent.

Confidence ratings were also collected to explore whether self-confidence contributed to reasons behind reliance on AI. Results indicate that confidence was predictive of changes in decision. Novices generally exhibited higher confidence and trust ratings than professionals, perhaps indicating that novices were less aware of their abilities than professionals. An explanation for this finding is that professionals and novices used the confidence scales differently. It was speculated that professionals would have previously received plenty of feedback on their face-matching performance during their training, and have a more accurate perception of their skills and abilities compared to novices. Novices who are relatively inexperienced have less exposure to such training and feedback and therefore may have approached the confidence scale differently by overestimating their

confidence. For both groups, AI assistance appeared to be related to confidence during the face-matching task, which led to similar trust and confidence ratings results. Results on the propensity to trust were in line with initial predictions, showing that individuals with a higher propensity to trust in general tended to display higher trust in the face-matching task. This indicates that individual differences in propensity to trust may play a role in how participants interact with AI systems.

The study showed that despite using AI with lower reliability, this did not undermine professionals' superior performance on face-matching tasks, as indicated by the results on sensitivity. The current sample of professionals consisted of facial examiners, facial reviewers and police investigators, a source of variability within professionals could be further explored to better define expertise. For instance, in fingerprint identification, the nature of expertise can be tested by manipulating the amount of information available, the opportunity for direct comparison, making use of memory or the time given to make an accurate decision (Thompson & Tangen, 2014). Whilst it was expected that professionals would have received formal training in face-matching to carry out their job, information on the type and duration of training could help better identify the aspects that make an individual a professional in the face-matching domain. Future studies could also define expertise by assessing either their perceptual skills or in terms of operational accuracy as these can be distinct concepts (Towler et al., 2018). An alternative experimental design where participants are given the option to not respond when they are not confident on a trial may also be better in capturing differences in face-matching approaches between professionals and novices.

Given that expertise can exist on a continuum (Sunday & Gauthier, 2018), simply comparing professionals and novices may not be sufficient. The current study recruited law enforcement professionals with the assumption that face-matching is an integral part of their job. However, various other occupations may also require participants to perform identity checks regularly, such as HR officers, or cashiers. Whilst the study ensured that no participants from the novices were facial examiners, facial reviewers and police investigators, no information on actual face-matching experience was gathered. While there were no known superrecognizers in the study, individuals with exceptional face-matching abilities may also be present in the sample. Finding ways to quantify expertise may be useful in future studies.

The original prediction was that due to novices' relative inexperience, they may rely more on AI assistance and place higher trust in the provided AI labels as a way to compensate for their lack of knowledge and expertise in the domain compared to professionals. Findings on confidence ratings were unable to confirm that utilisation of the AI was a result of comparisons between trust in the system and confidence in own abilities. Alternative explanations for the higher trust ratings in novices should be explored. For instance, answers in the debrief (see Appendix) suggest that novices may actually have limited exposure to the technology as they are not required to use it at work. Thus, it could be argued that novices had a more optimistic view of AI capabilities, leading to a higher level of trust in the system's recommendations.

Findings showed that confidence did not influence subsequent trust in AI but it was predictive of change outcomes, indicating that a decrease in confidence was associated with a change in decision after seeing AI advice. Accepting advice from the AI system despite that

some advice was inaccurate could be a result of avoiding complete rejection (Harvey & Fischer, 1997). An earlier noteworthy study examining reasons for taking advice found that participants take advice even from novice advisors who are less experienced and knowledgeable on a specific task. In the current study, participants demonstrated a tendency to refrain from outright rejection of the AI system despite recognising that the AI system has limited reliability. This explanation is plausible as there was no option not to use the AI.

Perhaps believing that the AI system was more capable than them, novices took the AI recommendation with the intention to improve. Results were somewhat similar to a previous study showing that the utilisation of an automated system is not based on relative confidence and trust values (Wiczorek & Meyer, 2019). The previous study described an explanation related to the perceived idea of teamwork and shared responsibility, where the allocation of functions between humans and AI was not well-defined. The current study has the limitation that how participants should use the AI is not clear. Future studies could explore how participants could integrate AI decisions with their own. Examining differences between professionals and novices may also be more apparent when examining tasks with higher stakes, as varying the risk involved could affect whether advice is taken from others (Wiczorek & Meyer, 2016).

Another factor that could be examined is the role of cognitive load. Complex tasks often increase cognitive load, which leads to omission errors (Lyell et al., 2018). Whilst the task was the same for both professionals and novices in the current study, the cognitive load experienced could have been different and this was not focused on in the current study. As novices were less familiar with the task of unfamiliar face matching, novices may have used

AI assistance as a way to reduce the cognitive burden and responsibility for mistakes. Results did not show evidence of this, as expertise was not a significant predictor of changes in decisions.

Tailoring AI support to the varying needs of individuals can enhance the overall performance and user experience, for example by taking into account their expertise and propensity to trust. The study was designed to explore the interaction between human face-matching decision-makers and facial recognition systems, with the aim to gain insight into the impact of expertise and face-matching with AI support on performance and trust, and eventually the optimization of human-AI interactions in real-world applications. The findings of the study contribute to our understanding of the relationship between expertise and AI support in unfamiliar face-matching, by confirming that professionals have higher sensitivity in general. The effect of label consistency was similar in both professionals and novices. There were no differences in bias, suggesting that unreliable AI assistance does not affect general tendencies to respond match or non-match in either professionals or novices. The study confirmed that trust ratings were generally higher for novices than professionals and results indicate that this difference was not an attempt by novices to compensate for a lack of confidence, or professionals exhibiting overconfidence in their abilities. To conclude, expertise appears to play a role in shaping trust and reliance on AI in face-matching tasks. Novices tended to have higher trust ratings in the AI system, which may suggest a great need for assistance and to validate their existing decisions, while professionals are more wary of system errors. Understanding differences in the trusting behaviours of professionals and novices can help optimize the design and implementation of AI systems in real-world face-matching applications.

Chapter 6: General Discussion

6.1 Introduction

The overarching aim of this research project was to explore the concept of trust calibration in human-AI collaboration, specifically focusing on unfamiliar face-matching. Matching unfamiliar faces is surprisingly error-prone, as demonstrated using face-matching tests which require individuals to review face pairs simultaneously to decide whether the faces belong to the same person or different people (Burton et al., 2010). Increasingly, AI-based technology including facial recognition systems is being used in the workplace, for example, in security settings and the use of e-gates. Face verification systems often include a human-in-the-loop, through a sequential set-up where humans are shown the outputs of AI before either accepting or rejecting the AI advice. However, research has shown that humans often get misled when face pairs are accompanied by labels such as 'same' or 'different' (Fysh & Bindemann, 2018), due to unconscious cognitive biases that shift uncertainty judgements (Howard et al., 2020). This suggests that the human-AI collaboration in a face-matching context is not reaching its potential, as fusing human and AI performance should increase accuracy further (O'Toole et al., 2007). The research project aims to verify the influence of using AI and explore trust calibration as a potential way to improve face-matching performance.

Trust calibration is a process that can help facilitate the interaction between humans and AI. When trust in AI is calibrated, the human operator can accept the output of a system when it is competent and make use of alternative resources or their own expertise when the system is unreliable (Lee & See, 2004). Trust is a multifaceted concept, influenced by variability in the human operating the system, environment or the system (Hoff & Bashir,

2015). Several research questions were developed with the aim of understanding the role of trust in the interaction between human face-matching decision-makers and facial recognition algorithms, influenced by factors such as AI reliability, propensity to trust, and professional expertise.

The first objective was to gain a better understanding of the impact of using AI in face-matching, looking at the accuracy of face-matching and uncovering any decision-making patterns when face-matching with AI (Chapter 3). The errors observed in human performance during face-matching are similar to findings in other human-AI collaborations such as fingerprint examiners and automated fingerprint identification systems (Dror & Mnookin, 2010), clinicians using clinical support systems (Goddard et al., 2014) and tasks assisted by automated decision aids such as baggage screening (Boskemper et al., 2022). Previous research on face-matching with AI had focused on one-to-many face-matching, for example, both experienced and inexperienced facial reviewers are susceptible to errors when presented with an increasingly longer list of candidate faces when making face-matching decisions (Heyer et al., 2018). The current research builds on previous research by examining the performance of an AI-assisted group in comparison to a group without AI support, with a particular focus on examining trust in these interactions. This exploration served as a first step to understanding the implication of AI in the specific domain of face-matching and facial recognition. Examining the effect of AI reliability on performance was also expected to provide insights into the practical applications of facial recognition technology, which might be particularly relevant in real-life settings where AI outputs may not be completely accurate. Human users of decision aids tend to be sensitive to different levels of reliabilities, as higher reliabilities tend to be related to higher agreement rates and higher confidence ratings (Wiegmann et al., 2001). In cases where optimal performance is

not achieved with highly reliable AI, then the focus is on the human as the limiting factor in the team collaboration. On the other hand, when AI is unreliable, then the human operator can use alternative resources to assist their decision-making. The interaction can be facilitated via trust calibration.

Further confirming the role of trust in this specific domain lays the foundation for further research on trust calibration. The third objective therefore was aimed to examine AI scores on face-matching performance (Chapter 4). It was designed to provide insights into the usefulness of alternative presentation formats of AI outputs. Presenting explanations whereby AI explains its recommendations, performance metrics and confidence information has been suggested to improve trust calibration (Zerilli et al., 2022). While prior research has demonstrated the effectiveness of fusing AI scores with human ratings to enhance face-matching performance (O'Toole et al., 2007), examining the presentation of AI scores as a standalone method has not been explored. In line with the theme of trust, the third objective also looked at AI scores as a tool to calibrate trust, as an additional AI system with the labels that are often presented to users.

The final objective recognised the relevance of expertise in the context of face-matching performance with AI (Chapter 5). While it is acknowledged that professionals generally have better performance compared to novices in face-matching, the impact of expertise in human-AI collaboration is an area that is worthy of research. Using automation is a trade-off between trust in the system and trust in the human operator's ability (Lee & Moray, 1992). Building on findings from other fields where different trusting behaviours of professionals and novices have been observed, the current project aimed to look at whether expertise could reduce bias when users are presented with AI support of low reliability. As a

preliminary hypothesis, the role of confidence was investigated as a mediating factor in the relationship between expertise, trust and performance. This final chapter aims to summarise the key findings and highlight the limitations of the research in general, along with recommendations for suggestions for future research. There will be a concluding remark at the end on face-matching with AI and trust calibration.

6.2 Summary of Findings

6.2.1 Key Findings: AI Reliability

The thesis focused on the use of AI in the face-matching decision-making process and all experiments detailed in the thesis involved using AI or presented labels that were similar to AI outputs. Pilot Study 1 first confirmed the impact of incorporating AI in a face-matching context. In the study, participants were provided face pairs from the KFMT and asked to decide whether they were faces of the same person or different people. Participants were allocated into four different groups, three of which involved using AI support and one group not using AI support. There were three variations of the AI support groups which involved presenting labels that were consistent or inconsistent with the face pair. In the reliable AI support group, participants were presented with fully accurate labels. In another group, participants were presented with some matched faces indicated as different while the second group had mismatched faces labelled as same.

Findings confirmed that using AI supports improved performance, indicated by a higher area under the curve (AUC) score in the reliable AI group compared to no AI support group. There were no differences between the other groups, suggesting that using AI with low reliability did not deteriorate performance more than not using AI support. Using AI with low reliability appeared to have a positive impact on human performance, though this

effect was not significant. Findings also showed that the influence of AI incorrectly advising ‘same’ on mismatch trials was similar to AI incorrectly advising ‘different’ on match trials, indicating that false alarms and misses of the AI had similar effects on human performance.

All subsequent experiments also involved using AI in a face-matching context and showed that introducing AI into face-matching influences decision-making. Experiment 1 examined the impact of AI reliability, by comparing the performance of three groups, using AI with high reliability, low reliability or without AI support. Unlike Pilot Study 1, Experiment 1 made use of the GFMT to first assess performance differences in the participants between the groups before introducing AI in face-matching. All participants in the experiment were asked to complete the GFMT without AI support and then allocated to an experimental group to complete the KFMT. In addition to providing a baseline for individual differences between the groups, comparisons were made between the two tests. Findings showed a more liberal response in the same participants, whether using AI with high or low reliability when matching faces with AI support from the KFMT. This effect was speculated to be more pronounced when AI has low reliability as comparisons between the groups in the KFMT showed similar findings. Introducing AI into the face-matching process introduced bias by increasing participants' tendency to respond ‘match’ rather than ‘non-match’ regardless of the truth.

These findings demonstrate the potential benefits of integrating AI into the face-matching process, as reliable AI appears to improve human performance to an extent that is better than not using AI support at all. However, issues may occur when AI has limited reliability by occasionally providing incorrect advice. Using AI, particularly AI with limited

reliability, increases response bias in the participants, which may increase the false alarm rates of the humans and this could have further implications in real-life contexts.

Trust in AI was examined in different experiments of the thesis. Experiment 1 showed that trust was higher in match trials than in non-match trials, which perhaps indicated that participants had higher trust in the AI's ability to match faces that are perceived to be similar. It could also related to own ability in face-matching, as accuracy tended to be higher for match trials than non-match trials. The results of Experiment 1 showed that AI reliability did not affect trust as trust ratings were similar across both AI support groups, using AI with high or low reliability. This could be a result of changes in trust due to different individual evaluations of AI reliability.

Label consistency had an interesting effect on trust, showing that trust ratings were higher in consistently labelled trials than inconsistently labelled trials. Experiment 2 in Chapter 4 showed that label consistency was a significant predictor of trust showing that trust was generally higher on consistently labelled trials than on inconsistently labelled trials despite not being provided feedback on whether the AI was correct or incorrect. Trust ratings were made solely on the AI labels and perceptions of the face pair. This suggested that participants perhaps developed an understanding of the AI's accuracy but this did not necessarily translate to better making better decisions. There was a dissociation between trust in the AI and performance as participants distrusted a label but chose to comply.

6.2.2 Key Findings: AI Dissimilarity Scores

Dissimilarity scores are a quantitative measure of the dissimilarity between two faces. The findings of Experiment 1 showed that combining normalised AI scores with human confidence ratings improved accuracy. Improvements were not higher than AI

performance alone, indicating that the human was still the limiting factor in the collaboration. Subsequently, Pilot Study 2 examined the impact of presenting dissimilarity scores to participants by asking them to match unfamiliar faces with or without AI support in the form of dissimilarity scores. Brief explanations clarifying the meaning of these scores were provided and all participants were asked to complete the test as accurately as possible, though there were no time limits. Sensitivity and bias were the primary indicators of performance in this experiment.

Findings showed that dissimilarity scores had an adverse impact on sensitivity, reducing participants' ability to discriminate between match and non-match face pairs. This suggested that displaying dissimilarity scores was not useful. Alternatively, the negative impact could be attributed to the AI's lack of accuracy, resulting in ambiguous scores that hindered rather than aided the decision-making process.

To address this, Experiment 2 used a more accurate face recognition model to generate dissimilarity scores. In the experiment, all participants were required to match faces using AI that provided inconsistent labels in around half of the trial, under two conditions, with or without AI dissimilarity scores. This experimental design aimed to examine the potential effectiveness of dissimilarity scores on trust calibration and examined whether dissimilarity scores interacted with label consistency to predict trust.

However, despite having used a more accurate facial recognition model, dissimilarity scores did not influence face-matching performance when paired with inconsistent labels. Resulted highlighted the prominent influence of labels, showing that dissimilarity scores no longer had an influence on face-matching decisions in the presence of AI labels. In line with findings from the previous experiment, results showed that the labels were predictive of

face-matching response, as participants responded non-match when the majority of faces were matched in the experiment and this was irrespective of whether the label was paired with AI dissimilarity scores or not.

These findings indicate that AI labels affect the ability to accurately match matched faces as participants relied heavily on the labels and likely overshadowed any potential impact of displaying AI dissimilarity scores. Further exploration of the effectiveness and interpretability dissimilarity scores may help understand this lack of influence. Understanding the interaction between decision-makers and AI dissimilarity scores may help address important questions on the optimal format of AI support in face-matching scenarios as changing the presentation of dissimilarity scores may be beneficial in optimising the interaction.

Experiment 2 also explored the role of propensity to trust automation and found that propensity to trust alone was not predictive of face-matching decisions. However, interaction effects showed that the effects of propensity to trust were significant in consistently labelled trials but not in inconsistently labelled trials. This finding showed that propensity to trust has potential value in the development of user-centred designs for facial recognition systems to optimise human-AI collaboration.

6.2.3 Key Findings: Face-Matching Expertise

Experiment 3 further confirmed that label predicted correct decision outcome, an effect that is evident in both professionals and novices. The experiment examined the effect of consistent and inconsistent labelling, as well as the role of trust, and its effect on professionals and novices by comparing their performance in the GFMT2, a version designed specifically for high-performing individuals. Police officers, facial reviewers and facial

examiners were recruited for the study as professionals and novices were age-matched controls. In the face-matching task, participants were required to first decide on the identity of a face pair, then presented with AI advice. Participants had the option to reconsider or change their initial response.

Results showed that in consistently labelled trials, professionals appear to have enhanced performance, however, similar to novices, their performance also deteriorated in inconsistently labelled trials. This shows the finding that AI labels influence face-matching decisions extends to professionals, indicating that in real-life contexts where identity verifications are carried out by face-matching professionals, even professionals may experience a challenge in maintaining performance when provided with AI that gives inaccurate advice.

Findings also showed that novices were more likely to change their face-matching decision after receiving AI advice. A plausible explanation could be linked to levels of confidence in their initial decision, as confidence ratings made before presented AI advice were predictive of changes in decision. These findings suggest that novices were more reliant on AI guidance and more likely to comply regardless of whether the advice given was accurate or inaccurate.

Experiment 3 further showed that there was no difference between professionals and novices in terms of trust ratings, and both professionals and novices were generally less trusting when the AI was showing an inaccurate label. Trust ratings were predicted by propensity to trust, the impact of which was significant for novices but not for professionals. This perhaps suggested that trust in AI by professionals was less likely to be related to their inherent propensity to trust and more related to AI reliability.

6.3 Contributions

The current research explored trust in AI in the context of face-matching and facial recognition systems, bringing insights from different fields into the psychology of face-matching and decision-making. The current thesis used experimental designs to validate the impact of AI on decision-making and aims to add practical value to contexts where human operator makers are assisted by AI, demonstrating its potential benefits of using reliable AI but also the potential bias and implications on trust when AI provides incorrect advice.

The experiments discussed in the thesis also used various approaches to examining face-matching performance. In addition to percentage accuracy, the thesis also used the signal detection theory to examine the effects on sensitivity and bias, as well as measures of performance using ROC/AUC. Taking into account the uniqueness of each face and potential issues with generalisability, mixed effects modelling was also used to account for random factors that may contribute to the effect of AI.

6.4 Implications

Research findings have important suggestions for the effective use of facial recognition technology in operational contexts. Using AI in face-matching introduces bias leading to higher false alarm rates of the human. The real-world consequences of inaccuracies in face-matching decisions could result in serious issues, including privacy concerns and legal implications. Increased false alarm rates could potentially waste valuable time and resources, leading to reduced efficiency in the workplace.

Current research findings have confirmed that the sequential set-up of human-AI collaboration where a human operator makes a decision after reviewing AI advice appears to be ineffective. Other collaborative options could be explored, for instance, understanding

task delegation dynamics may be useful (Fügener et al., 2022). This could inform how tasks should be distributed between humans and AI by considering both the human and AI's capabilities (Hemmer et al., 2023). This may address the limitations of simply presenting AI outputs and better optimise the collaboration.

Given that AI reliability influences performance outcomes, another option could be to prioritise the use of dependable AI systems with known rates of accuracy. This could be achieved by improving the facial recognition model or algorithm (Becerra-Riera et al., 2018), or standardising features of images that may affect accuracy such as low resolution, illumination, pose or facial expression. Places employing facial recognition technology could consider educating users about the limitations of facial recognition to reduce bias. This may enable users to make more informed decisions when face-matching with AI

Alternatively, there could be attempts to align users' trust levels with the actual reliability and performance of the AI system by exploring trust calibration methods. Current research findings confirmed the significant role of trust in the interaction between humans and facial recognition systems, varied by factors such as the type of trial, AI reliability and professional expertise. This has theoretical significance by contributing to the understanding of human-AI collaboration in the context of face-matching and highlighting the importance of trust in the relationship. Implementing measures against overtrust or under-trust in AI may be useful in addressing overreliance or under-reliance issues.

In terms of practical applications, specific methods for achieving trust calibration remain an open question. Presenting dissimilarity scores was not effective, though this could be further investigated to establish its interpretability to enhance usefulness. Developing trust calibration tools specifically designed for facial recognition may be a

promising area for future research. Research has explored the idea of explainable facial recognition using saliency maps (Lu et al., 2023). Another visualisation method could be the use of Grad-CAM, displaying regions of the faces that are emphasised in facial recognition models (Ito et al., 2021). Rebuilding trust after following instances of AI error may be valuable, especially as operators transition from using AI as a tool to perceiving AI to be a teammate (de Visser et al., 2018). These strategies may help mitigate the impact of AI providing inaccurate advice by improving the transparency of the AI.

Even when AI had limited reliability, professionals outperformed novices when given AI support. Both professionals and novices had reduced accuracy and trust in trials where AI provided inconsistent labels. These findings highlight both the need for reliable AI and the importance of expertise in face-matching. There was also a significant influence of propensity to trust in novices, suggesting that novices' inherent characteristic to trust may have guided their reliance on the AI.

Mechanisms behind expertise might be complex but potential solutions could include specifically recruiting professionals for face-matching tasks or developing specialist training and education to help individuals acquire expertise. Given the differences in the influence of propensity, the propensity to trust can be considered a valid measurement to be used in training and development and trust calibration strategies could be tailored to professional or novice and propensity to trust.

6.5 Limitations

6.5.1 Theoretical Considerations

Defining Trust.

Trust is an abstract concept that is challenging to define. This is because it is a complex construct of cognitive and affective processes (Lee & See, 2004), and varies due to a variety of different human-related or automation-related factors (Schaefer et al., 2016). Dispositions of the individual also have significant influences on shaping trust outcomes (Merritt & Ilgen, 2008).

Trust can be conceptualized as a psychological state. When referred to as an attitude, the perceived competence of a machine can be described as trust, which consists of elements such as predictability, dependability, and faith. This notion of trust is distinct from related concepts such as confidence, predictability, and accuracy (Muir, 1994). Due to the abstract nature of trust, it can only be inferred through behaviour.

Trust can be confused with confidence, which is a closely related concept. Vulnerability is what differentiates confidence from trust (Evans & Krueger, 2009). For instance, in our current experiments, unfamiliar face-matching is challenging enough to introduce uncertainty and can be argued to create a scenario where trust, rather than just confidence, is being measured. However, future studies could introduce negative or undesirable outcomes to further manipulate the level of vulnerability or risk, to clearly differentiate between trust and confidence.

Reliance is another closely related concept which has been used to infer trust. For example, previous research have shown that participants were more likely to rely on their own judgment rather than an automated system after witnessing the system make errors (Dzindolet et al., 2003). Trust-related behaviour might be a more favourable term to describe these interactions with a system, including measures of decision time, and compliance (Vereschak et al., 2021). Problems can arise when trust is not the sole

determinant of behaviour and is moderated by external factors such as risk (Satterfield et al., 2017). It is important to highlight that performance alone cannot be attributed to trust. Examining different trust-related behaviours could provide a more comprehensive understanding of trust in AI.

Understanding what trust is requires considering the broader context of how it is used, who uses it, and the societal context. For example, an international survey on public trust in AFR found that while there is general acceptance of its use, trust is higher when it is used by the police compared to the government and lowest in private companies (Ritchie et al., 2013). The study also found that Americans, in particular, are less comfortable with AFR than people in the UK and Australia. This indicates that trust varies depending on both the context and the entity employing the technology. Although our concept of trust focused specifically on unfamiliar face-matching, further defining the wider context could provide a more comprehensive understanding of how trust varies across populations and situations.

Measuring Trust.

The difficulty in defining trust contributes to the challenges of measuring trust accurately. Experiments in the current thesis used subjective ratings as measurements of trust but also recognised that it has limitations. Self-reports are frequently used to gather information on an individual's trust behaviour, attitudes, or intentions and there are many different forms of surveys or methods to measure trust such as behavioural outcomes or physiological measures (Kohn et al., 2021). Collecting subjective trust ratings has the advantage that it can easily be integrated into tasks and tends to have face validity. However, variations could arise due to factors like the type of scale used or the specific phrase of scale anchors. Whether focusing on a specific type or another measure of trust

yields different outcomes remains an unanswered question. Trust measurements rely on subjective data, which cannot always be verified. The nature of self-reporting introduces potential individual differences, as individuals may interpret and use rating scales differently. Experiments in the current research used continuous scales to maximise the precision of data.

The timing of trust measurement was less focused on, but it remains a relevant consideration. This was partly influenced by limitations related to the availability of facial stimuli. There were considerations on whether measuring trust before and after specific interactions with AI could yield informative insights. However, due to insufficient facial stimuli, it was difficult to ascertain whether there was sufficient time for participants to learn from interactions with the AI. Therefore, the primary emphasis was on examining trust in the moment of the interaction. Data was generally collected on every trial and may have caused some level of disruption, though this was expected to be minimal. Rest breaks were included in the experiment to address potential issues with fatigue or attention.

Defining Expertise.

The thesis focused on police officers, facial examiners and facial reviewers, which was also based on self-reported data, which relied primarily on participants' interpretation of the professional job title. Studying the role of expertise in a more naturalistic way using observations and qualitative methods as opposed to novice-expert experimental design may be an option to better understand the reasoning and decision-making process in a real context (Farrington-Darby & Wilson, 2006).

There were some attempts to quantify expertise by collecting data on the number of years the participant was in their occupation. However, it is recognised that this may not be

an accurate measure of expertise. Given that no one self-reported to be a superrecogniser, it can be assumed that expertise was acquired through on-the-job training. Particularly for the empirical study involving professionals, the availability of participants was limited. and therefore the definition of expertise was not made more specific. There were attempts to overcome this challenge by recruiting from professional groups to potentially increase the contrast between professionals and novices.

6.5.2 Real-life Contexts

Time Pressure.

It is recognised that the face-matching task used in these studies may be different to real-life face-matching scenarios. Time can significantly impact face-matching, for instance, time pressure can affect performance and potential biases are introduced over time (Fysh & Bindemann, 2017). Therefore, it is recognised that these studies only serve as a starting point for future research, where time constraints can be aligned more closely in practical settings. The current thesis focuses on performance measured by accuracy, sensitivity and bias, but it is also recognised that reaction time in conjunction with accuracy could be valid measures of performance. As experiments were online, there was naturally less control over task durations and time limits.

Purpose and Motivations.

It is also recognised that there are differences in motivations and incentives between the task in the current task and tasks in real life. Motivational incentives, such as food, improve face-matching accuracy, particularly for mismatched face pairs (Moore & Johnston, 2013). Participants in research studies may be aware of their role as research participants, whereas individuals in real-life contexts may view face-matching tasks as a part of their job

or responsibility. Real-life applications may have more severe consequences should mistakes arise, while research studies may only be motivated by monetary rewards, course credits or donation-based payments.

Availability of Faces.

A significant challenge encountered was the sourcing of appropriate face stimuli. The thesis focused specifically on face-matching, and required stimuli designed for this purpose. The number of trials in the studies of the current research was therefore constrained by the availability of faces in validated stimuli sets. In line with ethical standards, participants were informed about the number of trials and the estimated duration of the task. This differs from real-world scenarios, which may impact the generalisability of research outcomes.

6.5.3 Other Constraints

Facial recognition model.

The choice of facial recognition algorithm used in this research was primarily based on availability and convenience. However, it is acknowledged that facial recognition models and packages can offer a wide range of capabilities and variations in accuracy. For instance, the use of *face-recognition* in Chapter 3 and *Deepface* in Chapter 4 yielded different levels of accuracy when applied to faces from the KFMT. Using commercial facial recognition models may have led to different outcomes and findings. In some of the experiments in the current research, AI labels were manipulated to examine the effects of inaccurate advice with participants being informed that they were AI outputs. The current research findings are more broadly applicable to human-AI interaction in general rather than commenting specifically on particular algorithm models.

Online Experiments.

The nature of the online experiment involves technical considerations that could become limitations to the studies, including factors such as screen brightness and refresh rates. As face-matching can be a perception-based task, variations in colour and image-related inconsistencies could have influenced the results. Considering that internet speed and connectivity may also be an issue, the present study did not use reaction time as a performance measure, which could have given more comprehensive results. The technical abilities of participants were generally assumed, but it is recognised that familiarity with technology, and online experiments varies between participants. To address this issue, some of the studies in the current thesis included practice trials, though it is difficult to determine whether this was sufficient. There were attempts to review the quality of data after collection, excluding participants who did not meet the eligibility requirements or displayed signs of low attentiveness (Rodd, 2024). Given that familiar faces are processed differently than unfamiliar faces (Megreya & Burton, 2006), experiments in the research included trials with famous faces as attention checks.

Nature of Psychology Experiments.

A potential issue with most psychology experiments that may impact findings is that participants are naturally suspicious of such experiments. Participants may behave in ways that they believe will help the experiment, influenced by their attitude towards the study or the experimenter, particularly when they are aware of the study's hypothesis (Nichols & Maner, 2008). In all the current experiments in this thesis, regardless of whether a "real" algorithm was used, there was a possibility that participants doubted the authenticity of AI-generated decisions, despite being informed that AI was used.

The experiments in the current thesis are susceptible to this limitation, and the online nature of the studies naturally means less control over the experimental conditions. A potential way to address this issue is to provide participants with a misleading rationale or purpose for the study (Laney et al., 2007). For instance, by using a cover story and informing participants that the study is not about human-AI interaction but rather about a different topic. Additionally, participants' beliefs about the realism of the AI could be captured and assessed using post-experiment questionnaires.

6.6 Ethical Considerations

The current research acknowledges that facial recognition technology is associated with various ethical concerns. For example, the variability of accuracy across different demographic groups such as race and gender could lead to potentially unfair outcomes in applied settings (Abdurrahim et al., 2018). Given this concern, the current research aimed to improve face-matching accuracy by facilitating human-AI interaction. This reduces the need to rely solely on AI which brings a risk of bias and discrimination. Including a human-in-the-loop in the decision-making process adds additional responsibility to the decision-making process.

Another ethical concern is the privacy of individuals. The current research primarily used face databases that were for research purposes, where participants would have consented to have their face images taken. The current study used openly available facial recognition models, as opposed to proprietary software to examine face similarity. The current research emphasizes the responsible use of facial recognition technology and the need for greater transparency in AI systems, with a specific focus on one-to-one face matching in security applications.

6.7 Future Directions

The current research has shown a role for trust in explaining face-matching behaviour in the interaction between human decision-makers and facial recognition systems. Future research could investigate methods of calibrating trust in AI systems, specifically by examining the explainability of facial recognition. Enhancing AI transparency and making the process more comprehensible to users may help calibrate trust and address ethical concerns related to facial recognition technology. From a theoretical perspective, additional research could examine topics like trust violation and trust repair. As AI becomes more integrated into the workplace, a comprehensive understanding of the long-term impact of using AI could be useful and exploring strategies to repair the relationship with AI when errors occur may also help calibrate trust. Considering the potential inaccuracies of facial recognition algorithms across different demographics, further research could also investigate the interaction between human and algorithmic bias.

6.8 Conclusion

Face-matching is an important task as it is one of the most commonly used methods of identity verification, effectively making sure that faces are a match confirming the same identity or non-match, and preventing cases of fraud. In many applied contexts, human operators of facial recognition technology have to make a final decision if the system is unable to verify the identity. The current research examined trust calibration as a way to facilitate this process and a series of research questions were formulated, examining the influence of AI on face-matching, the impact of AI reliability, the presentation of AI dissimilarity scores, and the role of expertise. The research has highlighted the impact of using facial recognition technology, in particular the consequences of using AI with limited reliability. Implications of these findings were discussed, and while this research was unable

to verify ways to calibrate trust, research findings provide reassurance that trust calibration remains a possibility. This research serves as a stepping stone for future investigations in the field of psychology, face-matching and human-AI teams.

Appendices

Propensity to Trust Questions – Chapter 4

I believe that an Automated Facial Recognition (AFR) system is a competent performer.

"Strongly Disagree ", "Strongly Agree "

I trust an AFR system.

"Strongly Disagree ", "Strongly Agree "

I have confidence in the advice given by an AFR system.

"Strongly Disagree ", "Strongly Agree "

I can depend on an AFR system.

"Strongly Disagree ", "Strongly Agree "

I can rely on an AFR system to behave in consistent ways.

"Strongly Disagree ", "Strongly Agree "

I can rely on an AFR system to do its best every time I take its advice.

"Strongly Disagree ", "Strongly Agree "

Propensity to Trust Questions – Chapter 5

Generally, I trust AI.

"Strongly Disagree ", "Strongly Agree "

AI helps me solve many problems.

"Strongly Disagree ", "Strongly Agree "

I think it’s a good idea to rely on AI for help.

"Strongly Disagree ", "Strongly Agree "

I don’t trust the information I get from AI.

"Strongly Disagree ", "Strongly Agree "

AI is reliable.

"Strongly Disagree ", "Strongly Agree "

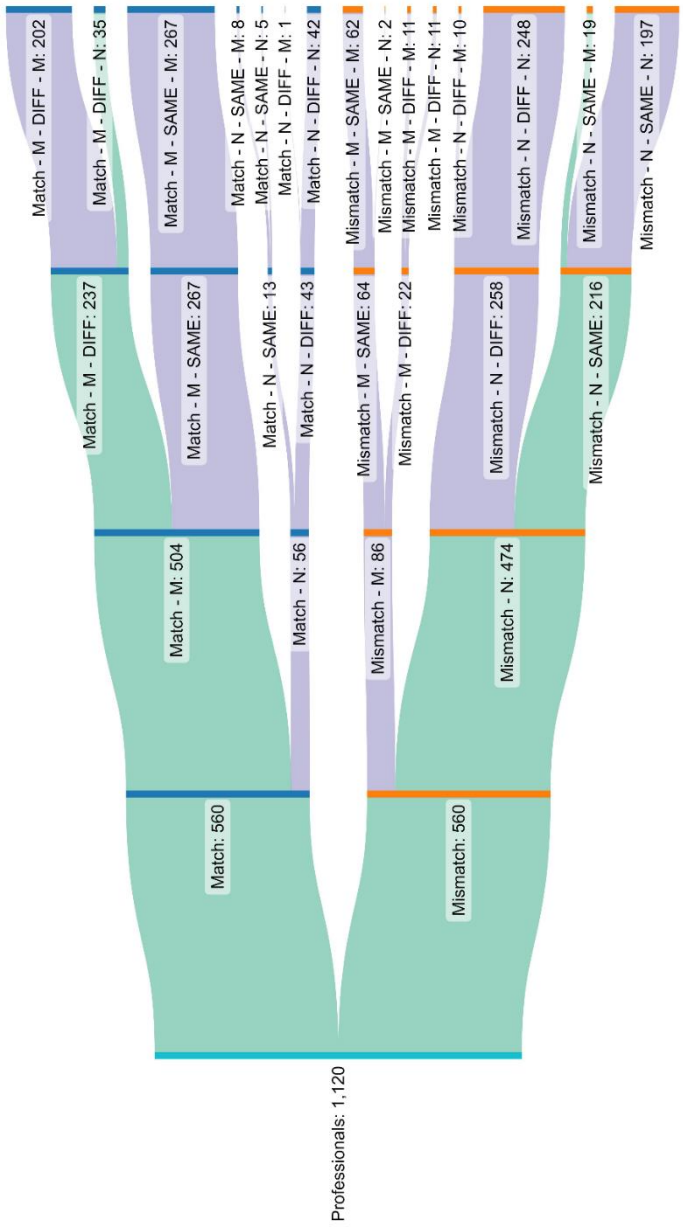
I rely on AI.

"Strongly Disagree ", "Strongly Agree "

Sankey Diagrams for Professionals – Chapter 5

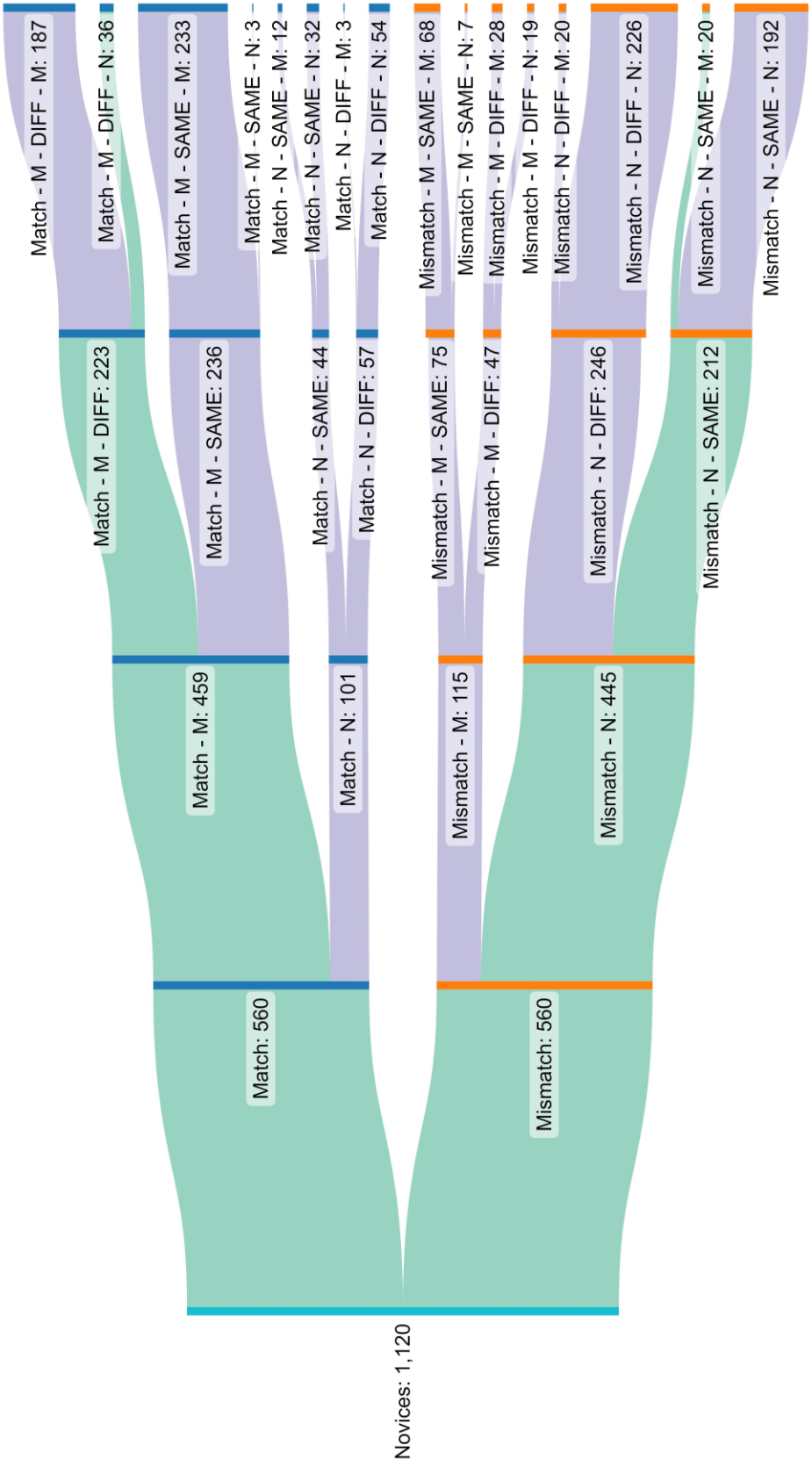
Diagrams showing responses of *match (M)* or *non-match* (N) on match and mismatch trials after being presented with the labels *same* (SAME) or *different (DIFF)*. Green flows indicate changes in response from a correct decision to an incorrect decision on inconsistently labelled trials.

Sankey diagram showing changes in decisions in professionals

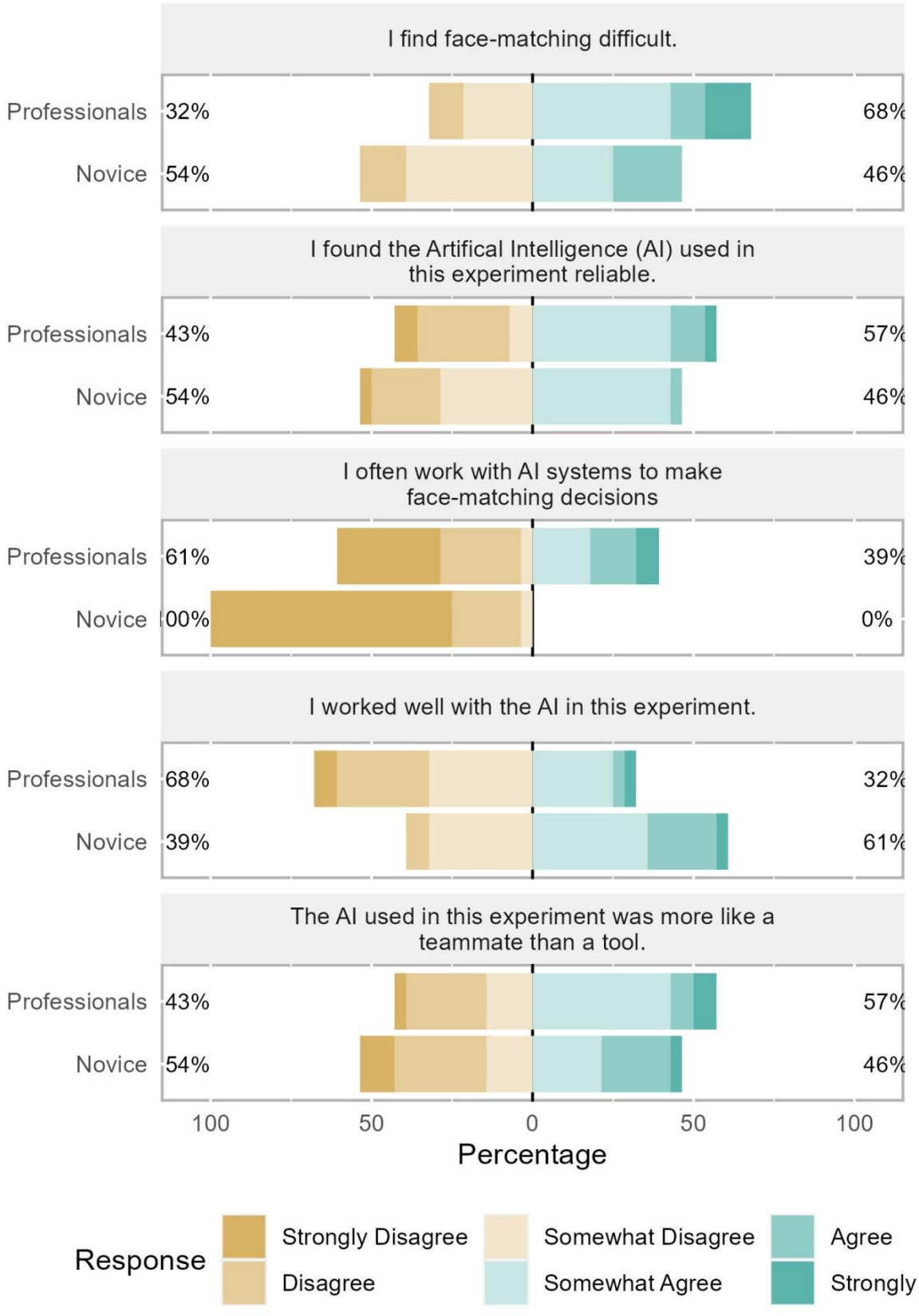


Sankey Diagrams for Novices – Chapter 5

Sankey diagram showing changes of decisions in Novices



Debriefing Questions – Chapter 5



References

Abdurrahim, S. H., Samad, S. A., & Huddin, A. B. (2018). Review on the effects of age, gender, and race demographics on automatic face recognition. *Visual Computer*, *34*(11), 1617–1630. https://doi.org/10.1007/s00371-017-1428-z

Agrawal, A. K., & Singh, Y. N. (2015). Evaluation of face recognition methods in unconstrained environments. *Procedia Computer Science*, *48*(C), 644–651. https://doi.org/10.1016/j.procs.2015.04.147

Albiero, V., Member, G. S., Zhang, K., King, M. C., Bowyer, K. W., & Fellow, L. (2022). Gendered Differences in Face Recognition Accuracy Explained by Hairstyles, Makeup, and Facial Morphology. *IEEE Transactions on Information Forensics and Security. 17*, 127–137. https://doi.org/10.1109/TIFS.2021.3135750

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

Alenezi, H. 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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bahner, J. E., Elepfandt, M. F., & Manzey, D. (2008). Misuse of Diagnostic Aids in Process Control: The Effects of Automation Misses on Complacency and Automation Bias.

Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 52(19), 1330-1334. https://doi.org/10.1177/154193120805201906

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust, *Theoretical Issues in Ergonomics Science, 8*(4), 321-348, https://doi.org/10.1080/14639220500535301

Baker, K. A., Stabile, V. J., & Mondloch, C. J. (2023). Stable individual differences in unfamiliar face identification: Evidence from simultaneous and sequential matching tasks. *Cognition*, *232*(February 2022), 105333. https://doi.org/10.1016/j.cognition.2022.105333

Becerra-Riera, F., Morales-González, A., & Méndez-Vázquez, H. (2018). Facial marks for improving face recognition. *Pattern Recognition Letters*, *113*, 3–9. https://doi.org/10.1016/j.patrec.2017.05.005

Beveridge, J. R., Givens, G. H., Phillips, P. J., & Draper, B. A. (2009). Factors that influence algorithm performance in the Face Recognition Grand Challenge. *Computer Vision and Image Understanding, 113*(6), 750–762. https://doi.org/10.1016/j.cviu.2008.12.007

Beveridge, J. R., Phillips, P. J., Givens, G. H., Draper, B. A., Teli, M. N., & Bolme, D. S. (2011). When high-quality face images match poorly. *IEEE International Conference on Automatic Face and Gesture Recognition (FG). 2011,* 572–578. https://doi.org/10.1109/FG.2011.5771460

Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition

accuracy for unfamiliar and famous faces. *Cogent Psychology*, *1*(1). https://doi.org/10.1080/23311908.2014.986903

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*(3), 277–291. https://doi.org/10.1037/a0029635

Bindemann, M., Fysh, M., Cross, K., & Watts, R. (2016). Matching Faces Against the Clock. *I-Perception*, *7*(5), 1-18. https://doi.org/10.1177/2041669516672219

Bobak, A., Mileva, V., & Hancock, P. (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

Bobak, A., 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). https://doi.org/10.1371/journal.pone.0148148

Boskemper, M. M., Bartlett, M. L., & McCarley, J. S. (2022). Measuring the Efficiency of Automation-Aided Performance in a Simulated Baggage Screening Task. *Human Factors*, *64*(6), 945–961. https://doi.org/10.1177/0018720820983632

Brewer, N. T., Gilkey, M. B., Lillie, S. E., Hesse, B. W., & Sheridan, S. L. (2012). Tables or bar graphs? presenting test results in electronic medical records. *Medical Decision Making*, *32*(4), 545–553. https://doi.org/10.1177/0272989X12441395

Brown, E., Deffenbacher, K., & Sturgill, William. (1977). Memory for Faces and the

Circumstances of Encounter. *Journal of Applied Psychology. 62*(3), 311-318. https://doi.org/10.1037/0021-9010.62.3.311

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *In Journal of Experimental Psychology: Applied, 7*(3), 207-218. https://doi.org/10.1037/1076-898X.7.3.207

Burton, M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence from Security Surveillance. Psychological Science, 10(3), 243-248. https://doi-org/10.1111/1467-9280.00144

Burton, M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behaviour Research Methods*, *42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Carbon, C. C. (2008). Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception*, *37*(5), 801–806. https://doi.org/10.1068/p5789

Carragher, D. J. (2023). Supplemental Material for Simulated Automated Facial Recognition Systems as Decision-Aids in Forensic Face Matching Tasks. *Journal of Experimental Psychology: General*, *152*(5), 1286–1304. https://doi.org/10.1037/xge0001310.supp

Castelfranchi, C., & Falcone, R. (2000). Trust is much more than subjective probability: mental components and sources of trust. *Proceedings of the Hawaii International Conference on System Sciences*, *May 2014*, 132. https://doi.org/10.1109/hicss.2000.926815

Cavazos, J. G., Jonathon Phillips, P., Castillo, C. D., & O'Toole, A. J. (2019). Accuracy

comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transaction Biometric, Behavior and Identity Science. 3*(1), 101-111. https://doi.org/10.1109/tbiom.2020.3027269

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2021). Accuracy Comparison across Face Recognition Algorithms: Where Are We on Measuring Race Bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *3*(1), 101–111. https://doi.org/10.1109/TBIOM.2020.3027269

Chavaillaz, A., Schwaninger, A., Michel, S., & Sauer, J. (2019). Expertise, automation and trust in X-ray screening of cabin baggage. *Frontiers in Psychology*, *10*(FEB), 1–11. https://doi.org/10.3389/fpsyg.2019.00256

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied Ergonomics*, *52*, 333–342. https://doi.org/10.1016/j.apergo.2015.07.012

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology*, *17*(1), 97–116. https://doi.org/10.1080/09541440340000439

Costa, A. C., Fulmer, C. A., & Anderson, N. R. (2018). Trust in work teams: An integrative review, multilevel model, and future directions. *Journal of Organizational Behavior, 39*(2), 169–184. https://doi.org/10.1002/job.2213

Davenport, R., & Bustamante, E. (2010). Effects of False-Alarm vs. Miss-Prone Automation and Likelihood Alarm Technology on Trust, Reliance, and Compliance in a Miss-Prone

Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(19), 1513–1517. https://doi.org/10.1177/154193120304701318

Davis, J., Atchley, A., Smitherman, H., Simon, H., & Tenhundfeld, N. (2020). Measuring Automation Bias and Complacency in an X-Ray Screening Task. *2020 Systems and Information Engineering Design Symposium, SIEDS 2020*, 6–10. https://doi.org/10.1109/SIEDS49339.2020.9106670

De Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012). The world is not enough: Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society*, *56*(1), 263–267. https://doi.org/10.1177/1071181312561062

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology, 106*(3), 433–445. https://doi.org/10.1111/bjop.12103

Dror, I. E., & Mnookin, J. L. (2010). The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. Law, Probability and Risk, 9(1), 47–67. https://doi.org/10.1093/lpr/mgp031

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role

of trust in automation reliance. *International Journal of Human Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors, 44*(1), 79-94. https://doi.org/10.1518/0018720024494856

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. Human Factors, 59(1), 5-27. https://doi.org/10.1177/0018720816681350

Evans, A. M., & Krueger, J. I. (2009). The psychology (and economics) of trust. *Social and Personality Psychology Compass, 3*(6), 1003–1017. https://doi.org/10.1111/j.1751-9004.2009.00232.x

Farrington-Darby, T., & Wilson, J. R. (2006). The nature of expertise: A review. *Applied Ergonomics*, *37*(1), 17–32. https://doi.org/10.1016/j.apergo.2005.09.001

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, *33*(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science, 26*(6), 797–815. https://doi.org/10.1016/S0364-0213(02)00084-8

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

018-0111-x

Fysh, M. C., & Bindemann, M. (2017). Effects of time pressure and time passage on face-matching accuracy. *Royal Society Open Science, 4, 170249,* http://doi.org/10.1098/rsos.170249

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

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology*, *109*(2), 219–231. https://doi.org/10.1111/bjop.12260

Graves, I. *et al.* (2011). The Role of the Human Operator in Image-Based Airport Security Technologies. In: Jain, L.C., Aidman, E.V., Abeynayake, C. (eds) Innovations in Defence Support Systems -2. Studies in Computational Intelligence, vol 338. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17764-4_5

Gill, H., Boies, K., Finegan, J. E., & McNally, J. (2005). Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust. *Journal of Business and Psychology*, *19*(3), 287–302. https://doi.org/10.1007/s10869-004-2229-8

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Godau, C., Vogelgesang, T., & Gaschler, R. (2016). Perception of bar graphs - A biased impression? *Computers in Human Behavior*, *59*, 67–73.

https://doi.org/10.1016/j.chb.2016.01.036

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*, *83*(5), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test (FRVT) part 3: Demographic effects*. National Institute of Standards and Technology.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Hagras, H. (2018). *Toward Human-Understandable, Explainable AI*. Computer, 51(9), 28-36. https://doi.org/10.1109/MC.2018.3620965

Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics*, *60*(2), 284–291. https://doi.org/10.1080/00140139.2016.1190035

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors, 53(5), 517-527. https://doi.org/10.1177/0018720811417254

Harvey, N., & Fischer, I. (1997). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2),

117–133. https://doi.org/10.1006/obhd.1997.2697

Hassaballah, M., & Aly, S. (2015). Face recognition: Challenges, achievements and future directions. *IET Computer Vision*, *9*(4), 614–626. https://doi.org/10.1049/iet-cvi.2014.0084

Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., & Satzger, G. (2023). Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 453–463. https://doi.org/10.1145/3581641.3584052

Herlitz, A. & Lovén, J. (2013) Sex differences and the own-gender bias in face recognition: A meta-analytic review, *Visual Cognition, 21*(9-10), 1306-1336 https://doi.org/10.1080/13506285.2013.823140

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), 1–18. https://doi.org/10.1371/journal.pone.0237855

Hu, Y., Jackson, K., Yates, A., White, D., Phillips, J, & O'Toole, A. (2017) Person recognition: Qualitative differences in how forensic face examiners and untrained people rely on the face versus the body for identification, *Visual Cognition, 25*(4-6), 492-506, https://doi.org/10.1080/13506285.2017.1297339

Hussein, A., Elsawah, S., & Abbass, H. A. (2020). The reliability and transparency bases of trust in human-swarm interaction: principles and implications. *Ergonomics*, *63*(9), 1116–1132. https://doi.org/10.1080/00140139.2020.1764112

Ingram, M., Moreton, R., Gancz, B., & Pollick, F. (2021). Calibrating trust toward an autonomous image classifier. *Technology, Mind, and Behavior*, *2*(1). https://doi.org/10.1037/tmb0000032

Ito, K., Kawai, H., & Aoki, T. (2021). A Comprehensive Study of Face Recognition Using Deep Learning. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 14-17*, 1762–1768.

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

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The Measurement of the Propensity to Trust Automation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11575 LNCS*. Springer International Publishing. https://doi.org/10.1007/978-3-030-21565-1_32

Johnston, R. A. & Edmonds, A. J. (2009) Familiar and unfamiliar face recognition: A review, Memory, 17:5, 577-596, https://doi.org/10.1080/09658210902976969

Jorritsma, W., Cnossen, F., & Van Ooijen, P. M. A. (2015). Improving the radiologist-CAD interaction: Designing for appropriate trust. *Clinical Radiology*, *70*(2), 115–122. https://doi.org/10.1016/j.crad.2014.09.017

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2017). Calibrating trust to increase the use of automated systems in a vehicle. *Advances in Intelligent Systems and Computing*, *484*, 535–546. https://doi.org/10.1007/978-3-319-41682-3_45

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, *12*, 1-23. https://doi.org/10.3389/fpsyg.2021.604977

Kose, N., & Dugelay, J. L. (2014). Mask spoofing in face recognition and countermeasures. *Image and Vision Computing, 32*(10), 779–789. https://doi.org/10.1016/j.imavis.2014.06.003

Kostka, G., & Meckel, M. (2021). Between security and convenience : Facial recognition technology in the eyes of citizens in China, Germany, the United Kingdom, and the United States. *Public Understanding of Science*, *30*(6), 671-690. https://doi.org/10.1177/09636625211001555

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*(1995), 569–598. https://doi.org/10.1146/annurev.psych.50.1.569

Kramer, R. S. S., and Ritchie, K. L. (2016) Disguising Superman: How Glasses Affect Unfamiliar Face Matching. *Applied Cognitive Psychology.* 30: 841–845. https://doi.org/10.1002/acp.3261

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors*, *62*(5), 718–736. https://doi.org/10.1177/0018720819853686

Lacson, F. C., Wiegmann, D. A., & Madhavan, P. (2005). Effects of attribute and goal framing on automation reliance and compliance. *Proceedings of the Human Factors and Ergonomics Society*, *49*(3), 482–486. https://doi.org/10.1177/154193120504900357

Lee, J. D., & Moray, N. (1994). Trust, Self-confidence, and operator's 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

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management*, *32*(6), 991–1022. https://doi.org/10.1177/0149206306294405

Lin, Y. S., Liu, Z. Y., Chen, Y. A., Wang, Y. S., Chang, Y. L., & Hsu, W. H. (2021). XCos: An explainable cosine metric for face verification task. *ACM Transactions on Multimedia Computing, Communications and Applications, 17*(3s). https://doi.org/10.1145/3469288

Lu, Y., Xu, Z., & Ebrahimi, T. (2023). *Towards Visual Saliency Explanations of Face Recognition*. 4726–4735. http://arxiv.org/abs/2305.08546

Lyell, D., Magrabi, F., & Coiera, E. (2018). The Effect of Cognitive Load and Task Complexity on Automation Bias in Electronic Prescribing. *Human Factors*, *60*(7), 1008–1021. https://doi.org/10.1177/0018720818781224

Lyons, J. B., Ho, N. T., Van Abel, A. L., Hoffmann, L. C., Sadler, G. G., Fergueson, W. E., Grigsby, M. A., & Wilkins, M. (2017). Comparing Trust in Auto-GCAS between Experienced and Novice Air Force Pilots. *Ergonomics in Design*, *25*(4), 4–9. https://doi.org/10.1177/1064804617716612

Ma, R., & Kaber, D. (2007). Effects of in-vehicle Navigation Assistance and performance on Driver Trust and Vehicle Control. *International Journal of Industrial Ergonomics, 37*(8), 665-673. https://doi.org/10.1016/j.ergon.2007.04.005

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*,*8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Madhavan, P., & Wiegmann, D. A. (2005). Effects of information source, pedigree, and reliability on operators' utilizaton of diagnostic advice. *Proceedings of the Human*

Factors and Ergonomics Society, 487–491. https://doi.org/10.1177/154193120504900358

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*(2), 241–256. https://doi.org/10.1518/001872006777724408

Makowski, D. (2018). The psycho Package: an Efficient and Publishing-Oriented Workflow for Psychological Science. *The Journal of Open Source Software*, *3*(22), 470. https://doi.org/10.21105/joss.00470

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Matthews, C. M., & Mondloch, C. J. (2018). Journal of Applied Research in Memory and Cognition Improving Identity Matching of Newly Encountered Faces: Effects of Multi-image Training. *Journal of Applied Research in Memory and Cognition, 7*(2), 280–290. https://doi.org/10.1016/j.jarmac.2017.10.005

Marinaccio, K., Kohn, S., Parasuraman, R., & De Visser, E. J. (2015). A Framework for Rebuilding Trust in Social Automation Across Health-Care Domains. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, 4*(1), 201-205. https://doi.org/10.1177/2327857915041036

Marsh, S., & Dibben, M.R. (2003), The role of trust in information science and technology.

Annual Review of Information Science and Technology, *37*(1), 465-498. https://doi.org/10.1002/aris.1440370111

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/AMR.1995.9508080335

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of Reliance and Compliance in Aided Visual Scanning. *Human Factors, 56*(5), 840-849. https://doi.org/10.1177/0018720813512865

McCaffery, J. M., & Burton, A. M. (2016) Passport Checks: Interactions Between Matching Faces and Biographical Details. *Applied Cognitive Psychology*, 30: 925–933. https://doi.org/10.1002/acp.3281

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*(4), 656–665. https://doi.org/10.1518/001872006779166334

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. PLoS ONE, 13(3), 1–16. https://doi.org/10.1371/journal.pone.0193455

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. (2008). Matching Faces to Photographs: Poor Performance in Eyewitness Memory (Without the Memory). *Journal of Experimental Psychology.*

14(4), 364–372. https://doi.org/10.1037/a0013464

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

Merritt, S. M. (2011). Affective processes in human-automation interactions. *Human Factors*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., Lachapell, J., & Lee, D. (2013). I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors*, *57*(1), 34–47. https://doi.org/10.1177/0018720814561675

Moore, R. M., & Johnston, R. A. (2013). Motivational incentives improve unfamiliar face matching accuracy. *Applied Cognitive Psychology*, *27*(6), 754–760. https://doi.org/10.1002/acp.2964

Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science*, *1*(4), 354–365. https://doi.org/10.1080/14639220052399159

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. https://doi.org/10.1080/00140139408964957

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human Computer Studies*, *45*(6), 669–678. https://doi.org/10.1006/ijhc.1996.0073

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology, 135*(2), 151–166. https://doi.org/10.3200/GENP.135.2.151-166

Norell, K., Läthén, K.B., Bergström, P., Rice, A., Natu, V. and O'Toole, A. (2015), The Effect of Image Quality and Forensic Expertise in Facial Image Comparisons. *Journal of Forensic Sciences, 60*(2), 331-340. https://doi.org/10.1111/1556-4029.12660

Oloyede, M. O., Hancke, G. P., & Myburgh, H. C. (2020). A review on face recognition systems: recent approaches and challenges. *Multimedia Tools and Applications, 79*(37–38), 27891–27922. https://doi.org/10.1007/s11042-020-09261

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLoS ONE, 15*(2), 1–20. https://doi.org/10.1371/journal.pone.0229132

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., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks. Trends in Cognitive Sciences,

22(9), 794–809. https://doi.org/10.1016/j.tics.2018.06.006

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*, *9*(4), 1–15. https://doi.org/10.1145/2355598.2355599

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. https://doi.org/10.1080/00140139.2012.691554

Papesh, M. H. (2018). Photo ID verification remains challenging despite years of practice. *Cognitive Research: Principles and Implications, 3,* Article 19. https://doi.org/10.1186/s41235-018-0110-y

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

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

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced complacency. *The International Journal of Aviation Psychology, 3*(1), 1–23. https://doi.org/10.1207/s15327108ijap0301_1

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

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160. https://doi.org/10.1518/155534308X284417

Phillips, P. J. (2011). Improving face recognition technology. *Computer, 44*(3), 84–86. https://doi.org/10.1109/MC.2011.87

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D., Dunlop, J., Lui, Y. M., Sahibzada, H., & Weimer, S. (2012). The good, the bad, and the ugly face challenge problem. *Image and Vision Computing, 30*(3), 177–185. https://doi.org/10.1016/j.imavis.2012.01.004

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., Chen, J. C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6171–6176. https://doi.org/10.1073/pnas.1721355115

Pintea, S., & Moldovan, R. (2009). The Receiver-Operating Characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies*, *9*(1), 49–66.

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors*, *57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rice, A., Phillips, P. J., Natu, V., An, X., & O'Toole, A. J. (2013). Unaware Person Recognition From the Body When Face Identification Fails. Psychological Science, 24(11), 2235-2243. https://doi.org/10.1177/0956797613492986

Ritchie, K. L., Cartledge, C., Growns, B., Yan, A., Wang, Y., Guo, K., Id, R. S. S. K., Edmond, G., Martire, K. A., Roque, M. S., & White, D. (2021). Public attitudes towards the use of automatic facial recognition technology in criminal justice systems around the world. *PloS One*, *16*(10), 1–28.

Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2017). Fraudulent ID using face morphs : Experiments on human and automatic recognition. *PLoS ONE*, *12*(3), 1–12. https://doi.org/10.1371/journal.pone.0173319

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PLoS ONE, 11*(2), 1–8. https://doi.org/10.1371/journal.pone.0150036

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, *134*, 104472. https://doi.org/10.1016/j.jml.2023.104472

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin and Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Salehi, P., Chiou, E. K., Mancenido, M., Mosallanezhad, A., Cohen, M. C., & Shah, A. (2021). Decision Deferral in a Human-AI Joint Face-Matching Task: Effects on Human Performance and Trust. *Proceedings of the Human Factors and Ergonomics Society*, *65*(1), 638–642. https://doi.org/10.1177/1071181321651157

Sanchez del Rio, J., Moctezuma, D., Conde, C., Martin de Diego, I., & Cabello, E. (2016). Automated border control e-gates and facial recognition systems. *Computers and Security*, *62*, 49–72. https://doi.org/10.1016/j.cose.2016.07.001

Sanchez, C., & Dunning, D. (2023). Are experts overconfident?: An interdisciplinary review. *Research in Organizational Behavior*, *43*(December), 100195. https://doi.org/10.1016/j.riob.2023.100195

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, *15*(2), 134–160. https://doi.org/10.1080/1463922X.2011.611269

Satterfield, K., Baldwin, C., de Visser, E., & Shaw, T. (2017). The influence of risky conditions on trust in autonomous systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 324–328. https://doi.org/10.1177/1541931213601562

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics*, *59*(6), 767–780. https://doi.org/10.1080/00140139.2015.1094577

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. *International Conference on Intelligent User Interfaces, Proceedings IUI*, *Part F1476*, 240–251. https://doi.org/10.1145/3301275.3302308

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A Unified Embedding for Face Recognition and Clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7-12*(2015), 815-823. https://doi.org/10.1109/CVPR.2015.7298682

Schwark, J., Dolgov, I., Graves, W., & Hor, D. (2010). The influence of perceived task difficulty and importance on automation use. *Proceedings of the Human Factors and Ergonomics Society*, *2*, 1503–1507. https://doi.org/10.1518/107118110X12829370088561

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers. 31*, 137–149. https://doi.org/10.3758/BF03207704

Streiner, D. L., & Cairney, J. (2013). Operating characteristics curves. *Springer Series in Reliability Engineering*, *59*(2), 187–207. https://doi.org/10.1007/978-1-4471-2966-0_10

Sunday, M. A., & Gauthier, I. (2018). *Face Expertise for Unfamiliar Faces: A commentary on Young and burton's "are we face experts"*. *Journal of Expertise, 1*(1). 35-41. https://www.journalofexpertise.org

Susa, K. J., Gause, C. A., & Dessenberger, S. J. (2019). Matching Faces to ID photos: The influence of Motivation on cross-race identification. Applied Psychology in Criminal Justice, *15*(1), 86-96.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science*, *182*(4116), 990–1000. https://doi.org/10.1126/science.182.4116.990

Tabererg, W. S. (2017). *The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education*. 48, 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Terhorst, P., Kolf, J. N., Huber, M., Kirchbuchner, F., Damer, N., Moreno, A. M., Fierrez, J., & Kuijper, A. (2021). A Comprehensive Study on Face Recognition Biases Beyond Demographics. *IEEE Transactions on Technology and Society*, *3*(1), 16–30. https://doi.org/10.1109/tts.2021.3111823

Thompson, M. B., & Tangen, J. M. (2014). The nature of expertise in fingerprint matching:

Experts can do a lot with a little. *PLoS ONE*, *9*(12), 1–23. https://doi.org/10.1371/journal.pone.0114759

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, *1*(4), 100049. https://doi.org/10.1016/j.patter.2020.100049

Towler, A., Kemp, R. I., Burton, M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PLoS ONE*, *14*(2), 1–17. https://doi.org/10.1371/journal.pone.0211037

Towler, A., White, D., Ballantyne, K., Searston, R. A., Martire, K. A., & Kemp, R. I. (2018). Are Forensic Scientists Experts? *Journal of Applied Research in Memory and Cognition*, *7*(2), 199–208. https://doi.org/10.1016/j.jarmac.2018.03.010

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating Training Methods for Facial Image Comparison: The Face Shape Strategy Does Not Work. *Perception, 43*(2-3), 214-218. https://doi.org/10.1068/p7676

Tredoux, C. (2002). A direct measure of facial similarity and its relation to human similarity perceptions. *Journal of Experimental Psychology: Applied*, *8*(3), 180–193. https://doi.org/10.1037/1076-898X.8.3.180

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine, 26*(8), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

Vasey, B., Ursprung, S., Beddoe, B., Taylor, E. H., Marlow, N., Bilbro, N., Watkinson, P., & McCulloch, P. (2021). Association of Clinician Diagnostic Performance with Machine Learning-Based Decision Support Systems: A Systematic Review. *JAMA Network Open*, *4*(3), 1–15. https://doi.org/10.1001/jamanetworkopen.2021.1276

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction, 5*(327), 1–39. https://doi.org/10.1145/3476068.

Von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy and Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

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). https://doi.org/10.1186/s41235-021-00288-z

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, *3*(4). https://doi.org/10.1098/rsos.150670

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), 1–14.

https://doi.org/10.1371/journal.pone.0139827

White, D., Guilbert, D., Varela, V. P. L., Jenkins, R., & Burton, A. M. (2021). GFMT2: A psychometric measure of face-matching ability. *Behaviour Research Methods*, *54*, 252–260 (2022). https://doi.org/10.3758/s13428-021-01638-x

White, D., Norell, K., Phillips, P.J., O'Toole, A.J. (2017). Human Factors in Forensic Face Identification. In: Tistarelli, M., Champod, C. (eds) Handbook of Biometrics for Forensic Science. Advances in Computer Vision and Pajttern Recognition. Springer, Cham. https://doi.org/10.1007/978-3-319-50673-9_9

White, D., Phillips, 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). https://doi.org/10.1098/rspb.2015.1292

White, D., Kemp, R. I., Jenkins, R., & Burton, M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin and Review*, *21*(1), 100–106. https://doi.org/10.3758/s13423-013-0475-3

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

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

Wiczorek, R., & Meyer, J. (2016). Asymmetric effects of false positive and false negative indications on the verification of alerts in different risk conditions. *Proceedings of the*

Human Factors and Ergonomics Society, 289–292. https://doi.org/10.1177/1541931213601066

Wiczorek, R., & Meyer, J. (2019). Effects of trust, self-confidence, and feedback on the use of decision automation. *Frontiers in Psychology*, *10*(MAR), 1–12. https://doi.org/10.3389/fpsyg.2019.00519

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*(4), 352–367. https://doi.org/10.1080/14639220110110306

Wilkinson, C., & Evans, R. (2009). Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science and Justice, 49*(3), 191–196. https://doi.org/10.1016/j.scijus.2008.10.011

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

Wu, H., Albiero, V., Krishnapriya, K. S., King, M. C., & Bowyer, K. W. (2023). Face Recognition Accuracy Across Demographics: Shining a Light Into the Problem. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2023-June*, 1041–1050. https://doi.org/10.1109/CVPRW59228.2023.00111

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating Effects of User Experience and System Transparency on Trust in Automation. *ACM/IEEE International Conference on Human-Robot Interaction*, *Part F1271*, 408–416.

https://doi.org/10.1145/2909824.3020230

Young, A., & Burton, M. (2018). Are We Face Experts?. *Trends in Cognitive Sciences. 22*(2), 100-110. https://doi.org/10.1016/j.tics.2017.11.007

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *3*(4), 100455. https://doi.org/10.1016/j.patter.2022.100455

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852