EVALUATING FACE IDENTIFICATION EXPERTISE: TURNING THEORY INTO BEST PRACTICE

Digested Analysis

August 2020

Workshop sponsored by the Academy of the Social Sciences in Australia, UNSW Sydney School of Psychology and the UNSW Scientia Fellowship scheme.

Location: AGSM Building, UNSW Sydney, Australia
Dates: Monday 6 & Tuesday 7 January 2020

Convenors: David White, Alice Towler, Linda Jeffery, Richard Kemp & Romina Palermo.
Discussion Leaders and Speakers: Kaye Ballantyne, Kim Curby, Gary Edmond, Kristy Martire, Alice O'Toole,Jonathon Phillips, Mehera San Roque & Jeremy Wilmer.
Workshop Members: Thomas Carter, James Dunn, Cameron Tullberg, Daniel Ferguson, Jeannine Geach, Rebecca Heyer, Dana Michalski, Reuben Moreton, Eilidh Noyes, Kay Ritchie & Clare Sutherland

Please direct correspondence to:
Dr. David White | david.white@unsw.edu.au

PREFACE

The workshop described here aimed to provide evidence-based guidance on what it means to be an 'expert' in face identification. Together with Artificial Intelligence tools (AI), experts in this task play essential roles in safeguarding the accuracy of critical face identification decisions made in a variety of settings from courts to national borders.

This workshop was timely for two reasons. First, scientific knowledge of face identification expertise has grown substantially in recent years. Second, a sharp increase in the use of facial biometrics is expected to result from the Identity Matching Services Bill[1], which is currently before the Parliament of Australia. Resultant changes to national identity management procedures are expected to require a new face identification workforce to perform manual adjudication of facial image comparison decisions and oversee the functioning of national scale face identification systems.

The workshop was convened by cognitive psychologists David White, Romina Palermo, Linda Jeffery, Richard Kemp and Alice Towler. Our scientific backgrounds span applied and theoretical approaches to understanding human performance in face identification tasks. We have also worked with practitioners and interdisciplinary teams of legal, technical and policy experts in addressing practical issues relating to face identification in legal and security settings.

It was from these collaborative projects that the need for a common definition of expertise in face identification became clear. To meet this need, we assembled a core team of cognitive psychologists, computer scientists and legal scholars, each with extensive experience in face identification research. We also invited police and government employees with backgrounds in practice and policy relating to face identification and identity management.

We thank all the workshop members for their thoughtful, generous and energetic input over the two days of the workshop and look forward to working together to disseminate the outcomes in the months and years ahead. The digested analysis was initially drafted by the convenors, followed by rounds of critical review by the discussion leaders and speakers, and finally by all workshop members. A list of workshop members can be found here.

We invite feedback on this document from academics, practitioners and industry stakeholders. Please direct comments to david.white@unsw.edu.au.

Dr. David White
UNSW Sydney, August 2020

[1] Identity-matching Services Bill 2019.

CONTENTS

EXECUTIVE SUMMARY

Critical face identification decisions that underpin security, forensic and legal processes are often made by people. Border control officers compare a passport photo to a traveller, surveillance officers see a person of interest in a crowd, police officers compare mugshots to CCTV images. Nowadays, many of these tasks are supported by AI facial recognition technology[2]. Policymakers, academics and the general public are debating how this technology should be used, and the appropriate privacy and human rights safeguards.

An important point that is often overlooked when deploying face recognition technology is that it does not fully automate face identification decisions. People are integral to these decisions because human oversight can ensure accuracy, accountability and ethical use.

Face identification decisions can have negative impacts on people's lives, potentially restricting their access to government services, freedom to travel across national borders or even leading to their wrongful arrest[3,4]. Face identification systems, which incorporate AI and human decision-making, can be designed to limit these negative impacts, and ensure that they do not disproportionately affect socio-economic or demographic groups.

To address these emerging issues, we convened an international workshop of researchers in face identification from psychology, forensic science, artificial intelligence and law — with practitioners and policy-makers from police and government (see Workshop Members). We hope that outcomes can assist in development of policy and implementation of face identification and identity management systems in government, police, private industry and the judicial system. The main conclusions and recommendations of the workshop are:

- ***Face identification is now a mature multi-disciplinary field incorporating forensic science, cognitive psychology and artificial intelligence research.*** Compared to other biometric and pattern-matching disciplines, there is extensive research on the performance of humans and face recognition technology in face identification tasks. This research provides a foundation of scientific understanding that can provide the basis for designing accurate, fair, responsible and transparent human use of face recognition technology.

- ***Recent research shows that accuracy of the best Artificial Intelligence (AI) face recognition technology and the best humans are comparable, but performance is optimized by combining decisions made by the best AI and the best humans.*** A key challenge is to incorporate these research findings into operational systems with appropriate human oversight. To do this, it is first necessary to have agreed protocols

[2] Centre for Data Ethics and Innovation (2020) Snapshot series: Facial recognition technology report.

[3] Wrongfully accused by an algorithm (2020). New York Times.

[4] Georgetown Law (2015). The perpetual line up: Unregulated police face recognition in America; Georgetown Law (2019). Face recognition on flawed data.

for determining what are the 'best' performing people and face recognition technologies. We refer to best-performing solutions as face identification 'experts'.

- ***Face identification 'experts' must consistently demonstrate superior performance on tasks representative of the claimed expertise.*** The workshop unanimously agreed that qualification as an 'expert' in making face identification decisions should be based solely on proven superior performance — not on secondary indicators of expertise like a person's professional experience or training. Experts can be trained staff, novices with natural talent in the task or indeed AI technology, so long as their superior performance has been demonstrated. This definition can help create an effective face identification workforce, guide better design of face identification systems and provide the basis for legal definitions of expertise that are used to determine the admissibility of expert testimony in court.
- ***There is substantial variation in accuracy and performance between individual experts, and between different face recognition algorithms.*** Research shows variable accuracy amongst even the most accurate algorithms and humans. Patterns of errors also vary depending on the type of face identification decisions being made. For example, certain people and algorithms make more errors on faces from certain demographic groups. Progress is being made in creating calibrated tests for human and algorithm performance that can help select appropriate experts for specific tasks and reduce bias in face identification systems.
- ***New types of expert practitioners and researchers are required to design, evaluate, oversee, and explain modern face identification systems. Because these systems incorporate human and AI decision making, people with broad expertise in related disciplines are required.*** The workshop members are part of the emerging field of *face identification*, which is characterised by an integration of applied and theoretical questions, and of research and practice. Multidisciplinarity of our field entails that: (i) the next generation of researchers should be 'multilingual' in the discipline areas that intersect this new field; (ii) future face identification *practitioners* will require more diverse knowledge of forensic science, psychology and artificial intelligence to use face recognition technology appropriately; (iii) organisations deploying face identification systems will require similarly diverse expertise to implement, manage, evaluate, and explain these complex systems.

Part 1 of this report provides background to the workshop. Part 2 is a digested analysis of our discussion, outcomes and recommendations. Part 3 captures discussions on future research directions, which are primarily directed towards researchers in this field. In this section, we also outline plans for disseminating workshop outcomes and sustaining collaboration between academics, policy-makers and practitioners. A detailed record of the meeting schedule is provided in Appendix A1.

PART 1: WORKSHOP BACKGROUND

1.1 WHY IS FACE IDENTIFICATION IMPORTANT?

Our faces are the primary means of identifying one another, both when recognising people we know in everyday life, and in important applied settings (e.g. border control, police investigations and criminal trials). Despite improvements in other types of biometric technology over the past century, large volumes of identity checks are still made by people, comparing images of unfamiliar faces and deciding whether they are the same person or different people. This task, which we refer to as *face matching*, was the primary focus of the workshop.

Recent advances in facial recognition technology, coupled with the ever-increasing number of cameras that capture facial images via personal devices and CCTV, has led to a large increase in the volume of digital images captured for identity authentication and used in criminal investigations and trials. This means that there is, in fact, a *growing* need for humans to perform face matching tasks in end-to-end face identification systems – where “face identification system” refers to an organisation’s complete process of producing an identification decision from start to finish. Society therefore requires professionals with expertise in face identification more than ever before, and new types of expert with the appropriate skills, knowledge, and training to oversee these large-scale face identification systems.

1.2 CURRENT AND FUTURE CHALLENGES IN FACE IDENTIFICATION

Despite improvements in the accuracy of facial recognition technology, large volumes of face identification decisions will continue to be made by humans for the foreseeable future. Counter-intuitively, increases in biometric matching capability resulting from the widescale deployment of facial recognition technology typically *increase* the need for manual identity resolution of face image pairs. This increased workload arises because automation enables face matching to be performed at a far greater scale than ever before, leading to new identity resolution tasks. For example, where police use facial recognition technology to search mugshot databases using images from CCTV or social media[5], or immigration officers use the technology to cross-check visa applications[6].

Human oversight of facial recognition technology is necessary. As will be seen in the remainder of this report, end-to-end face identification systems that incorporate AI and

[5] Georgetown Law (2015), footnote 4; Davies, B., Innes, M., & Dawson, A. (2018). An evaluation of South Wales police’s use of automated facial recognition. Universities’ Police Science Institute Crime and Security Research Institute, Cardiff University.

[6] White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PloS one*, *10*(10), e0139827; Noyes, E. & Hill, M, Q. (in press). Automatic Recognition Systems and Human Computer Interaction in Face Matching. In M. Bindemann (Ed.), Forensic face matching: Research and practice: Oxford University Press.

human decision making can improve accuracy compared to facial recognition technology alone. Of equal importance, humans provide legal safeguards to protect against errors made by automated technology and promote ethical use of AI. But this need for human oversight is also problematic because many decades of research has shown that the average person is surprisingly poor at matching the identity of unfamiliar faces. To demonstrate the difficulty of this task, the reader is invited to guess how many people are pictured in Figure 1.

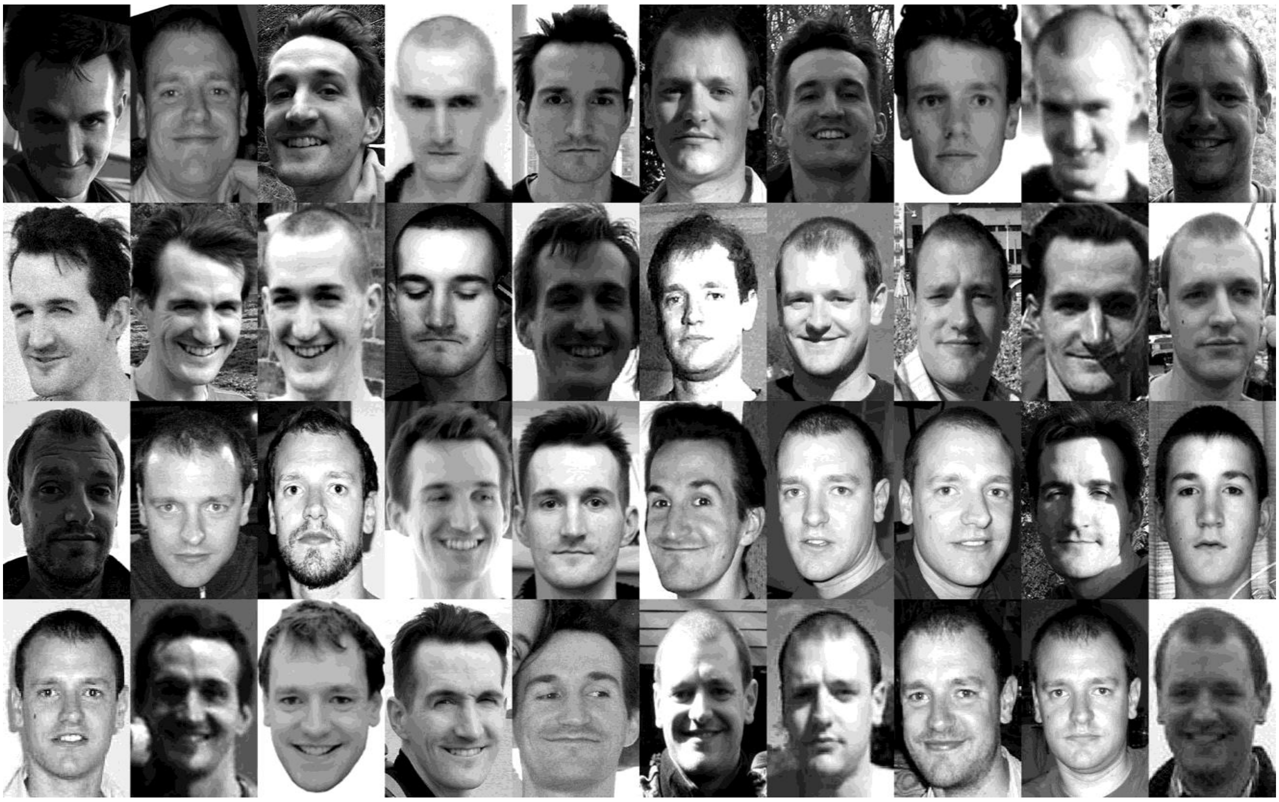

FIGURE 1. How many people are in this set of images? See the text below and Appendix A2 for the answer. Images sourced from Jenkins et al. (2011)[7].

Most people find this a difficult task. When psychologist Dr. Rob Jenkins asked novice participants how many people appeared in this array, the average participant answered that there were 7 people[8]. In fact, there are just 2 people in this array (see Appendix A2 for the solution). Importantly, this task is only difficult when we are *unfamiliar* with the faces. When we show this array to colleagues of the people pictured, they all answer correctly. Our ease in accurately recognising the faces of those familiar to us may explain why many people mistakenly expect to be accurate in recognising unfamiliar faces[9].

Poor human performance in unfamiliar face matching is a key challenge to address when designing face identification processes. In applied settings where staff make face

[7] Jenkins R, White D, Van Montfort X, Burton AM (2011) Variability in photos of the same face. *Cognition*, 121, 313-323.

[8] see footnote 7.

[9] Ritchie KL, Smith FG, Jenkins R, Bindemann M, White D, & Burton AM (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. Cognition, 141, 161-169; Young, A. W., & Burton, A. M. (2018). Are we face experts?. Trends in cognitive sciences, 22(2), 100-110.

identification decisions as part of their work, the faces are almost always unfamiliar to them. Recent research examining performance of professional staff shows that in many groups, performance is just as error-prone as in studies of novice participants. For example, when asked to match pairs of faces such as those shown in Figure 2, both university students and passport issuance officers made *1 error in every 5 decisions* [10].

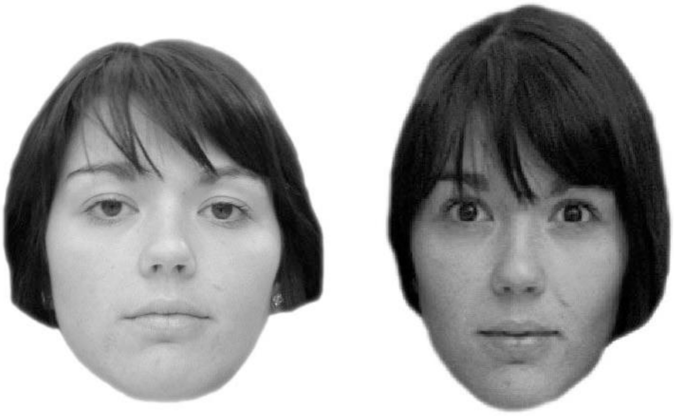

FIGURE 2. An example test item from the Glasgow Face Matching Test[11]. This task requires participants to decide whether pairs of images show the same person or different people. Despite the images being taken on the same day, with neutral expression, standardised pose and lighting conditions, both novice participants and passport issuance officers make 20% errors on this task.

1.3 TYPES OF FACE IDENTIFICATION PERFORMED BY HUMANS

"Face identification" is an umbrella term that encapsulates all tasks where a decision is made about the identity of a face. Broadly speaking, there are two classes of face identification task in applied settings.

Matching-based tasks require comparison of face images and a decision regarding whether they show the same or different people. The tasks shown in Figure 1 and Figure 2 are examples of matching tasks that have been developed for assessing face matching performance in the lab. Outside the lab, these tasks are performed in the daily work of professionals such as border control officers, forensic examiners and government employees that issue national identity documents (e.g. passports).

In contrast, *memory*-based tasks involve recognising a face as one that has been seen previously. Often, in applied memory-based tasks, the officer or witness is not personally familiar with the person of interest and may have only seen them in an image or from a fleeting glance. This makes the task much harder than recognising faces that we know well. For example, when a surveillance officer is searching for a person of interest in a crowded space, or when an eyewitness picks out a suspect from a police line-up. Compared to matching tasks, errors on memory tasks for unfamiliar faces are even more frequent. Meta-

[10] White D, Kemp RI, Jenkins R, Matheson M, Burton AM (2014) Passport officers' errors in face matching. *Plos One*, 9(8), e103510.
[11] Burton AM, White D, McNeill A (2010) The Glasgow Face Matching Test, Behavior Research Methods, 42(1), 286-291.

analysis of many decades of research on human performance in police line-up tasks shows that, even under optimal conditions, people make errors on 50% of cases[12].

Different professional roles require different face identification tasks. For example, a surveillance officer may require good memory for faces while a border control officer or forensic examiner will require strong matching ability. Studies of novice populations show that a person's ability in face memory tasks is highly correlated with ability in face matching tasks, but expertise in one of these tasks does not guarantee expertise in the other[13].

1.4 WHAT IS A FACE IDENTIFICATION EXPERT?

Scientific consensus after decades of research is that an expert is someone who demonstrates *superior performance* on a task that is representative of their domain of expertise. In society more broadly, for example in the legal system and workplaces, other criteria are often applied to assess expertise. Our aim was to derive a definition that was grounded in the available empirical evidence, so we focused on scientific studies of human performance to inform our discussions and the recommendations contained in this report.

Although the average person is likely to make many errors when identifying unfamiliar faces, there is also growing evidence that certain groups achieve high levels of accuracy[14]. In addition, facial recognition technology can now achieve levels of accuracy that exceed the accuracy of the average human. Further studies have shown advantages of combining decisions of humans with those of facial recognition algorithms[15]. Consequently, there are people, algorithms, and combinations of human and algorithms — which we term '*Hybrid' human-AI expert systems* — that can be considered to be 'experts' in face identification.

HUMAN EXPERTS

Two main groups of professional experts have been identified in empirical tests of face identification accuracy. First, facial examiners who have extensive professional training and experience in face matching tasks. Second, super-recognisers working for police forces, who have been selected based on their natural aptitude for making accurate face identification decisions. When tested on the task shown in Figure 2, facial examiners and super-recognisers both achieve substantially higher accuracy than novices.

[12] Steblay N, Dysart J, Fulero S, Lindsay RCL (2001) Eyewitness accuracy rates in sequential and simultaneous lineup Presentations: A Meta-Analytic Comparison. Law Hum Behav 25, 459-473.

[13] Verhallen RJ, Bosten JM, Goodbourn PT, Lawrance-Owen AJ, Bargary G & Mollon JD (2017). General and specific factors in the processing of faces. Vision Res 141, 217-227; Bate S, Frowd CD, Bennetts R, Hasshim N, Murray E, Bobak AK, Wills H & Richards S (2018). Applied screening tests for the detection of superior face recognition. Cognitive Research: Principles and Implications, 3, 1-19.

[14] Phillips PJ, Yates AN, Hu Y, Hahn CA, Noyes E, Jackson K, Cavazos JG, Jeckeln G, Ranjan R, Sankaranarayanan S, Chen JC, Castillo CD, Chellappa R, White D, O'Toole AJ (2018) Face recognition accuracy of forensic examiners, superrecognisers, and face recognition algorithms. PNAS, 115(24), 6171-6176.

[15] O'Toole AJ, Abdi H, Jiang F, Phillips PJ (2007). Fusing face-verification algorithms and humans. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 37(5), 1149-1155.

Facial examiners are specialists in matching-based tasks and their role requires them to produce detailed, analytic forensic reports of similarities and differences in facial and image features for criminal investigations and trials. Facial examiners appear to have acquired their expertise through professional training and experience, and their performance is characterised by conservative responding and relatively slow response times[16].

On the other hand, super-recognisers appear to have acquired their expertise over the course of their everyday lives[17]. Certain police forces have recruited officers specifically for their natural ability in face memory and face matching tasks. Typically, these officers do not receive extensive training or mentorship before being deployed in operational roles. Although current deployment of super-recognisers is limited to police and private investigation, there are reports of super-recognisers providing testimony in court, and there are many roles where they could potentially improve face identification in applied settings[18].

EXPERT ARTFICIAL INTELLIGENCE (AI)

The past decade has seen rapid improvements in AI and facial recognition technology through the use of 'deep learning' neural networks. In the most recent tests, state-of-the-art algorithms outperform the average human in face matching tasks and are comparable to the very best human experts[19]. This means that facial recognition technology available on the open market is now more accurate than the average human at matching unfamiliar faces. However, it should also be noted that some current deployments in operation do not use the most accurate technology available on the market.

Given this rapid advance, it is tempting to conclude that AI will soon replace human processing. This is very likely in many applied settings. For example, at border control, the processing burden of face identification decisions has gradually shifted from humans to AI over the past decade. However, the type of 1-to-1 matching between a passport image and an image taken by an automated border gate represents only one of the many deployments of this technology in modern society. In many other applications, hybrid systems that incorporate human and machine processing are more prevalent and will continue to be for the foreseeable future[20].

HYBRID HUMAN-AI EXPERT SYSTEMS

Improvements in facial recognition algorithms have enabled new applications of this technology, many of which have either created new identification tasks or changed the nature of tasks that people are required to perform. An overview of the two main ways humans interact with facial recognition technology is illustrated in Figure 3.

[16] Towler A, Kemp RI, White D (in press). Can face identification ability be trained? In M. Bindemann (Ed.), Forensic face matching: Research and practice: Oxford University Press.

[17] see footnote 16.

[18] Forensic Science Regulator (2018). Annual Report November 2016– November 2017.

[19] see footnote 14.

[20] Towler A, Kemp RI, White D (2017) Unfamiliar face matching systems in applied settings. In M Bindemann & AM Megreya (Eds.), Face Processing: Systems, disorders and cultural differences. New York: Nova Science Publishers Inc.

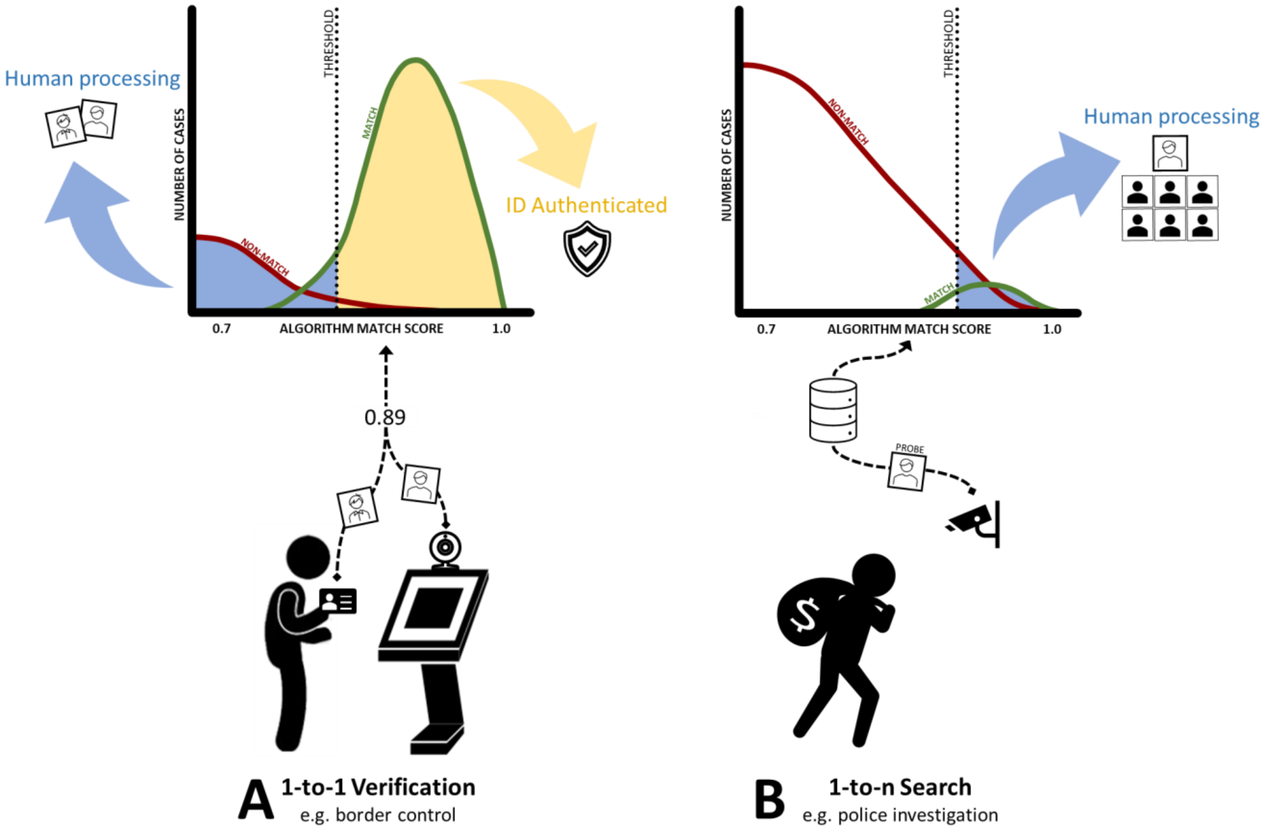

FIGURE 3. Two examples of hybrid human-AI expert systems, where humans interact with facial recognition technology. (A) 1-to-1 verification systems like those used at border control divert cases that do not meet a set threshold of similarity for human processing. Thresholds are set that minimize risk of imposters gaining access while being sensitive to the capacity for human processing given traveller volumes and staff numbers. Nevertheless, overlap of frequency distributions of matching faces (green frequency curve) and non-matching faces (red curve) mean that a proportion require manual processing by border control staff. (B) 1-to-n database search systems like those used in police investigation return ‘candidate lists’ of faces for human adjudication that exceed a threshold of similarity to a ‘probe’ image. Thresholds are set that determine the size of these candidate lists, but because database images with highest match scores contain both matching and non-matching faces, human adjudication is necessary.

For example, in “1-to-1 verification” applications – including automated border gates – facial recognition software is used to verify that faces match (see Figure 3A). The algorithm will assess the similarity between the two images – known as the algorithm *match score*. Where the system is unable to verify the match, because the match score falls below a set *threshold*, the case will be referred for human processing. A primary line officer will then decide whether the traveller matches the image on their passport, or whether they should be sent for secondary processing by immigration agents. This is an example of the requirement for humans to work alongside AI to ensure proper oversight.

Because there is an overlap in the match scores that are generated by matching faces (green frequency curve in Figure 3A) and non-matching faces (red frequency curve), some travellers with genuine passports will be diverted along with imposters carrying false documentation. The degree of overlap between the match (green) and non-match (red) distributions and the resulting volume of human processing can be reduced by improving the algorithm or image capture conditions. However, some overlap is unavoidable – e.g. due to changes in a persons' appearance over time, and lookalikes. Also, human processing is sometimes necessary for other reasons, such as when the automated gate at border control is unable to capture a suitable image of the face (e.g. baldness can cause issues), or to satisfy an organisation's procedural and legal requirements.

Another example of how humans interact with facial recognition technology is in police investigations, where staff will use it to perform a "1-to-n search" for an unknown offender, using a 'probe' image, in a large database of known identities (see Figure 3B). Often, these searches are performed on databases containing millions of images, and image quality of probe and database images is typically more variable than in the passport verification example above. As a result, high match scores returned by the algorithm contain a mixture of matching and non-matching faces, as denoted by the overlapping green and red frequency curves in Figure 3B. Because this type of image search was not possible before facial recognition technology was developed, the deployment of this technology has introduced a new workload for humans, who must adjudicate 'candidate lists' of database images that exceed the system threshold. Importantly, this 1-to-n application is becoming more and more prevalent. For example, to protect against identity fraud when issuing identity documents[21], to identify suspects in criminal investigations[22] and to monitor video and image data streams for 'watchlists' of target identities[23].

As the effectiveness of facial recognition technology improves, the frequency distributions of match scores for matching and non-matching identities – the green and red distributions in Figure 3 – move further apart, and so less human processing is necessary. However, an equally important implication of this trend is that the cases which *do* require human processing *will become increasingly more challenging*. These changes in the number and type of cases requiring human processing may require humans within these systems to have different kinds of face identification expertise. For example, consider the case of identical twins. It is feasible that the humans adjudicating future systems will need specialist knowledge and skills that enable them to discriminate between twins – something which algorithms find challenging. Similarly, it might be necessary for human experts to have a detailed understanding of how algorithms might get decisions wrong, so they can identify cases where that is likely, and take appropriate steps to avoid an error.

[21] see White et al.(2015), footnote 6.
[22] Georgetown Law (2015), footnote 4.
[23] Davies, B., Innes, M., & Dawson, A. (2018). An evaluation of South Wales police's use of automated facial recognition. Universities' Police Science Institute Crime and Security Research Institute, Cardiff University.

1.5 WHY DO WE NEED A DEFINITION OF EXPERTISE IN FACE IDENTIFICATION?

Definitions of expertise guide important real-world decisions. When human resource departments hire staff, judges decide whether to admit expert evidence in court, or when educators design training pathways, they do so by considering what constitutes “expertise”. Defining what it means to be a face identification expert can therefore provide a critical basis for designing staff recruitment procedures, developing training, and assessing the value of expert evidence in legal settings. Currently, there is no consensus on a definition of expertise in face identification.

The lack of a definition of face identification expertise has direct and serious real-world implications. For example, courts do not have clear guidance on who, or indeed *what* should be considered a face identification expert. Establishing expertise is central to determining which evidence is admissible at trial, and without evidence-based guidance on what constitutes expertise there is a risk that poor quality identification evidence may be admitted *and that high-quality identification evidence is unnecessarily excluded*. At present, both scenarios are common in Australian courts, but this is also a problem worldwide, as the UK Forensic Science Regulator recently outlined[24].

There is also increasing demand for face matching practitioners. In Australia for example, the enactment of the Identity-matching Services Bill (2019)[25] will lead to a capability for face matching decisions to be assisted by facial recognition technology on a national scale — to verify identity in transactions between citizens and government/industry, and to enable new capabilities in identifying suspects in criminal investigations. Similar national scale systems are being implemented in many other countries[26]. Because people are required to manage, analyse and adjudicate these systems, this will transform the type of expertise required in governments across the world. To meet this need, a face identification workforce with the necessary skills to perform the increasingly challenging face identification decisions that will arise from increased use of facial recognition technology is required. Without a definition of face identification expertise, it is not clear how these experts should be recruited or trained.

1.6 THE WORKSHOP

The purpose of the workshop was to develop an evidence-based definition of face identification expertise, and to develop evidence-based guidelines for conceptualising and assessing expertise in face identification.

In designing the workshop, we attempted to facilitate broad discussions that incorporated the different types of human and AI ‘experts’ described in Section 1.4, as well as the different types of tasks that are performed in professional settings. We also aimed to anticipate how

[24] Dodd V (2020) Forensic science failures putting justice at risk, says regulator. The Guardian, 25 February.

[25] see footnote 1.

[26] Home Office. (2018). Biometrics Strategy Better public services Maintaining public trust.

these roles and tasks may change in the future, particularly due to advances in facial recognition technology.

Defining expertise is complex because high levels of face identification accuracy can be achieved by diverse and heterogeneous types of face expert– i.e., humans, algorithms, and hybrid human-AI systems. Even taking human performance alone, there are qualitatively different types of expertise supporting high levels of accuracy in facial examiners compared to super-recognisers.

The workshop provided the unique opportunity to address this challenging and complex problem by bringing together world-leaders in face identification research with practitioners, policy-makers and legal scholars (see Workshop Members).

PART 2: WORKSHOP DISCUSSIONS, OUTCOMES AND RECOMMENDATIONS

In this section, we summarise the main points of consensus that emerged from the workshop discussions. These fall into four broad themes: A definition of expertise in face identification, Evaluating expertise in face identification, Key scientific findings, and Considerations for designing end-to-end face identification systems.

2.1 A DEFINITION OF EXPERTISE IN FACE IDENTIFICATION

Having a definition of expertise can facilitate the creation of an effective face identification workforce, help design better face identification systems, and provide the basis for legal definitions used to determine admissibility of expert testimony in court. However current definitions do not incorporate up-to-date scientific understanding of expertise in face identification.

For example, legal definitions of expertise that are used to determine admissibility of face identification evidence are too general, so fail to reflect advances in the science and technology of face identification. Section 79 of the Uniform Evidence Act states that an expert is someone who has "specialised knowledge" based on their "training, study or experience". This definition precludes people with natural skill in face identification from providing face identification evidence in court, despite compelling scientific evidence they perform more accurately than many people with training, study or experience (see Key Finding 1)[27].

To address the limitations of current definitions, the workshop aimed to reach consensus on a definition of face identification expertise that is consistent with current scientific understanding of expertise in face identification and accepted scientific definitions of expertise more broadly[28]. The workshop members agreed on the following definition of expertise, which can be applied to the full gamut of face identification tasks as well as the different types of experts described in Section 1.4:

> ***"Expertise is the consistent demonstration of superior performance on task(s) representative of the claimed expertise."***

Key aspects of this definition, and those that distinguish it from traditional and currently

[27] White D, Towler A, Kemp RI (in press). Understanding professional expertise in unfamiliar face matching. In M. Bindemann (Ed.), Forensic face matching: Research and practice: Oxford University Press.

[28] Ericsson KA & Lehmann AC (1996) Expert and exceptional performance: Evidence of maximal adaptation to task constraints. Annual Review of Psychology 47, 273-305.

accepted definitions are as follows:

The core requirement of experts is a "consistent demonstration of superior performance". Surprisingly, empirical tests of ability are typically not required to be considered an expert in face identification. A demonstration of superior performance means having the ability to present performance data on tests that are representative of the task of claimed expertise[29] (see Testing Principle 1), where performance is demonstrably higher than an appropriate benchmark (see Testing Principle 3). For example, someone who claims to be able to verify the identity of travellers at border control must demonstrate that when they compare the faces of travellers to passport photographs they can discriminate between matching and non-matching identities. At a minimum, they should be able to match faces with a superior level of accuracy compared to the average person. However, appropriate performance benchmarks also depend on the particular applied context, and so should be set by taking into account the acceptable chances of error in a given context.

We deliberately include the term "consistent" in our definition to acknowledge that high performance on a small number of test cases can be achieved simply by chance, and because the title of "expert" should be reserved for those who possess a skill that persists over time. Superior performance must therefore be demonstrated multiple times in order to be a reliable indicator of someone's true ability. Cognitive psychologists, who are trained in behavioural research methods, are well-placed to determine the number and configuration of tests that would be required to demonstrate consistent superior performance in face identification experts (see Testing Principle 2).

Experts are defined by specifying the "claimed expertise". Claimed expertise refers to the task or skill that a person (or technology) claims to perform with a high degree of accuracy. For example, a border control agent would claim they can quickly determine whether a passport photo belongs to the traveller or a different person. A surveillance officer might claim they can memorise a face and recognise the person again in a crowd several hours later. A forensic facial examiner might claim they can decide if simultaneously presented CCTV images show the same person or different people. These practitioners might also claim they can accurately communicate their reasoning and analysis processes, visual comparison techniques, and the strength of evidence to judges and jurors. To determine the claim to expertise, practitioners should ask themselves "What is it that I claim to be able to do?". Practitioners may have many claims to expertise.

Importantly, expertise in one face identification task (e.g. border control) does not necessarily transfer to others (e.g. identifying people from CCTV footage), because different skills underlie each task. Superior performance must therefore be demonstrated for *every* claim to expertise. There is no need to be an expert in a particular face identification task if

[29] Some practitioners and policy-makers may feel that routine proficiency tests meet this criteria. However, most proficiency tests are entirely inadequate for assessing expertise. Most critical is that proficiency tests are typically not representative of casework because they are much easier and/or do not mimic the typical task requirements, and because many use consensus marking rather than accuracy.

the role does not involve that task. Thus, if an expert does not claim to have expertise in a particular face identification task, there is no need to demonstrate superior performance.

Training and/or professional experience – no matter how extensive or prestigious – is insufficient evidence to support claims of expertise. We deliberately did not include training and experience in our definition of expertise because there is no direct evidence that training or experience either improves accuracy, or contributes to the development of expertise in face identification tasks. Simple before-and-after tests used to evaluate training courses in facial image comparison show little evidence of improvement[30], and there are no studies that have tracked the development of professional expertise over time. In addition, there was agreement that face identification abilities that are based on natural talent, particularly those relating to face memory tasks, are not likely to be 'trainable' beyond an individual's inherent capacity, given existing evidence showing training is largely ineffective in producing generalised improvement on these tasks[31]. In time, research may reveal that training and experience are important for developing face identification expertise[32], but until this has been established empirically they should not be considered evidence of expertise in face identification.

Similarly, seniority, anecdotal reports, endorsements by colleagues, and subjective judgments of competence are meaningless in determining expertise. The science provides clear evidence that these metrics bear no relationship to face identification expertise. For example, accuracy is not related to practitioners' years of experience, and people have poor insight into their own face identification abilities[33], and the effectiveness of training[34].

The definition of expertise is inclusive of the diverse types of face identification expertise. As described in Section 1.4, there are different types of experts that all achieve high levels of accuracy. For example, many individuals who have no experience performing professional face identification tasks are nonetheless extremely accurate, by virtue of their *natural ability* (super-recognisers). Our definition captures all sources of face identification expertise, regardless of whether it originates from training, experience, or natural aptitude and is agnostic to whether the expert is human, an algorithm, or a hybrid human-AI system–so long as consistently superior performance is demonstrated via performance on representative tests.

[30] Towler A, Kemp RI, Burton AM, Dunn JD, Wayne T, Moreton R, White D (2019) Do professional facial image comparison training courses work? Plos One, 14(2), e0211037.

[31] see footnote 16.

[32] Given the proven superior accuracy of forensic facial examiners, it is likely that some of their training and/or professional experience is effective in improving accuracy on facial image comparison tasks. Studies also suggest that this improvement is related to enhanced ability to compare individual facial features. However, formal testing of the effectiveness of training, mentorship and internship programs are required for these to be validated as pathways to expertise in face identification tasks.

[33] Zhou X, & Jenkins R (2020) Dunning-Kruger effects in face perception. Cognition, 203, 104345; Bindemann M, Attard J & Johnston RA (2014) Perceived ability and actual recognition accuracy for unfamiliar and famous faces. Cogent Psychology, 1, 986903; Bobak AK, Mileva VR, & Hancock PJB (2019) Facing the facts: Naïve participants have only moderate insight into their face recognition and face perception abilities. Quarterly Journal of Experimental Psychology, 72(4), 872-88134.

[34] see footnote 30.

2.2 EVALUATING EXPERTISE IN FACE IDENTIFICATION

To demonstrate expertise, the workshop members agreed that face identification experts must show consistently superior accuracy on tests that are representative of the task they are claiming to have expertise in (see Section 2.1).

Establishing expertise therefore requires tests that provide reliable measures of accuracy in the various face identification tasks. Although there are already tests that meet these criteria for some face identification tasks, it will be necessary to create new tests that enable sustainable evaluation of expertise across the many different face identification tasks, and at the scale that will be required if our recommendations are widely adopted. To facilitate this test development work, the workshop members agreed on four main principles that should be followed when creating tests to evaluate face identification expertise. These principles can be applied to the development of bespoke tests, to be used within organisations to examine accuracy on specific tasks, and also to standard tests that are more broadly applicable and can enable standardised performance benchmarks to be established.

TESTING PRINCIPLE 1. TESTS EVALUATING EXPERTISE MUST BE REPRESENTATIVE OF THE TASK THE EXPERT IS CLAIMING TO HAVE EXPERTISE IN

Applied face identification tasks vary enormously. Some involve simultaneous comparison of images, whereas others involve recognition of memorised faces. Some involve comparison of studio-quality images, while others involve poor-quality CCTV images. Tasks may be assisted by face recognition technology, performed on images presented on computer screens, or performed in live settings (for example comparing a person to a photo-ID). There is a long list of other factors that contribute to the diversity of applied face identification tasks including: the demographic makeup of the faces that are being identified, whether faces are being matched across variations in age, the frequency with which non-matching pairs are expected to be encountered, and the time that can be spent reaching an identification decision.

Given this diversity, there is no one-size-fits-all solution to establishing expertise in face identification. Instead, tests must be designed to measure performance on the precise task of claimed expertise. For example, a practitioner who matches CCTV stills to mugshots should have their expertise assessed differently to a practitioner who matches passport photos, despite the basic format of the task being identical (i.e. 1-to-1 comparison). In research, this quality of a test is known as *test validity*: the extent to which a test measures the ability it was designed to measure[35].

[35] Duchaine B & Nakayama K (2006) The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. Neuropsychologia 44(4), 576-585; Bowles DC, McKone E, Dawel A, Duchaine B, Palermo R, Schmalzl L, Rivolta D, Wilson CE & Yovel G (2009) Diagnosing prosopagnosia: Effects of ageing, sex and

The first step in designing a valid test is to gain a thorough understanding of the face identification task in question. This initial step can be surprisingly complicated, and this challenge was relayed to the group by workshop members who are currently working towards this goal in a forensic science context[36]. Even apparently straightforward tasks often involve multiple sub-tasks. For example, before forensic facial examiners match faces, they typically assess whether the images are of sufficient quality to enable this judgement. Examiners' ability to perform this screening sub-task directly affects their ability to perform the face identification task accurately, because it prevents costly errors made based on poor-quality evidence. Tests should therefore be careful to capture all aspects of the claimed expertise. If need be, tests should measure components separately where appropriate.

There are numerous factors to consider when designing a test to assess a claim to expertise. Many of these are covered in Martire and Kemp's (2016)[37] general guidance for assessing human performance in applied settings. Below we outline some important additional factors to consider when assessing expertise in face identification tasks specifically. For example, it is important that the face images used to construct a test are representative of the faces that an expert will be required to identify:

- ***Is there a requirement to identify faces of different ethnicities?*** Many years of face recognition research shows that people are poorer at identifying faces from an ethnic group other than their own[38]. Therefore, expertise that is verified based on tests containing White European faces may not extend to faces of other ethnicities, and vice versa[39]. Test creation should therefore be sensitive to the demographic makeup of the faces that an expert is required to identify.

- ***Are the images of children?*** Children's faces are subject to substantially more variation in appearance over time than adult faces, and accuracy on face identification tasks that include children's faces are typically much lower than equivalent tests of adult faces[40]. It is therefore possible that expertise in identifying children's faces is qualitatively different to expertise in identifying other faces, and so this should be reflected in tests of expertise where the task is to identify children's faces.

participant-stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. Cognitive Neuropsychology, 26(5), 423-455.

[36] The National Institute of Forensic Sciences are currently carrying out the "Fundamentals of Forensics Project" to systematically map each forensic science discipline's claims to expertise. This work is critical for validating forensic science procedures and developing robust industry proficiency tests.

[37] Martire KA, Kemp RI (2016) Considerations when designing human performance tests in the forensic sciences. Australian Journal of Forensic Sciences, 50(2), 166-182.

[38] Meissner CA, Brigham JC (2001) Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. Psychology Public Policy and Law, 7(1), 3-35.

[39] Bate S, Bennetts R, Hasshim N, Portch E, Murray E, Burns E, & Dudfield G (2019) The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills. Journal of Experimental Psychology: Human Perception and Performance, 45(3), 363.

[40] Michalski, D. (2017). The impact of age-ralated variables on facial comparisons with images of children: Algorithm and practitioner performance (Doctoral Dissertation). University of Adelaide, Australia.

- ***Is there a requirement to identify faces across changes in age?*** Face matching accuracy decreases as the time between each face image being captured increases[41]. Therefore, if experts wish to claim expertise in tasks that involve matching across large differences in age, then this type of decision should be incorporated into the test.
- ***What is the typical quality and resolution of the imagery?*** The identity information contained in images changes depending on the quality and resolution of the imagery. Identifying faces in different image quality conditions therefore requires different skills. For example, fine-grained analysis of facial features is thought to be a critical source of identity information for facial examination, but poor-quality imagery will often not contain the fine feature detail necessary to permit this type of comparison. Claims to expertise must therefore be tested on imagery that reflects the particular face identification task.

Task properties should also be represented in tests that assess expertise, for example:

- ***Does the task involve matching, memory or both?*** While the correlation between accuracy on face matching and face memory tasks is relatively high in novices, indicating these tasks recruit common cognitive mechanisms, there is also evidence that they are not the same abilities. Moreover, in cases where expertise has been acquired primarily through professional experience, for example in forensic facial examiners, the expertise is likely specific to face matching. Therefore, it is critical that claims to expertise in face memory and matching tasks are validated using different tests.
- ***Are cues from body, clothing or other contextual cues available?*** In many face identification tasks there are other cues available[42]. For example, a CCTV image will often capture the body and clothing of the person, not just their face. Depending on the claim to expertise that is being made, it may be appropriate to create tests that contain these other identity cues.

The challenge in making tests that are representative of real-world tasks is two-fold. First, researchers must know what tasks experts perform in their daily work. Second, they must design tests that best capture the essential properties of these real-world tasks. At present, the first part is a major hurdle, because there is limited crosstalk between practitioners and academics.

Recent collaborative work between researchers and industry is a promising first step to resolving this issue, but it is essential that this program continues to expand so that experienced scientists can develop tests that are useful in applied settings. Once tests have been created, they must be available to practitioners and their employers so they have the tools necessary to assess claims to expertise. Where claims to expertise are unsupported,

[41] Michalski D, Heyer R, Semmler C (2019) The performance of practitioners conducting facial comparisons on images of children across age. PLoS ONE 14(11): e0225298

[42] Noyes E, Hill MQ & O'Toole AJ (2018) Face recognition ability does not predict person identification performance: Using individual data in the interpretation of group results. Cognitive research: principles and implications, 3(1), 1-13.

training and recruitment strategies may be able to be adjusted appropriately. Development of representative tests is therefore dependent on a continuous cycle of knowledge development, '*from the lab to the world and back again*'[43].

TESTING PRINCIPLE 2. TESTS EVALUATING EXPERTISE MUST BE RELIABLE AND STABLE INDICATORS OF ABILITY

While Testing Principle 1 is concerned with test validity, an equally important consideration is test *reliability*. Test reliability is the likelihood that the test will provide the same result if repeated, and so is an indicator of how stable ability is over time. Reliability is represented in our definition of expertise by "*consistent* demonstration of superior performance".

Reliability is an important property of tests of expertise because the expert will continue to make important decisions for some time after completing a test. We must therefore have some confidence that the test provides a reasonably accurate indication of future performance.

Both test reliability and test validity are fundamental properties of psychometric testing and are studied extensively by behavioural and cognitive scientists. Some existing tests of face identification ability show high levels of reliability which shows that face identification ability is something that can be measured accurately. These tests are created by people with a background and training in psychometric approaches and behavioural research methods. To ensure high reliability and desirable psychometric properties, tests should be created by or in consultation with behavioural scientists with the relevant training.

A final point is that repeatedly performing the *same* test is problematic. This is because people are likely to learn the answers to the test or become familiar with the faces contained in the test. People's performance may therefore increase on each repeat for reasons unrelated to their skills, which means the test no longer provides a measure of a person's expertise. At present, only three standard tests of face identification ability exist that are suitable for testing experts. A critical aim for future work is therefore to create tests that contain multiple equivalent versions that can be used to track expertise over time. Such tests will ensure experts can reliably demonstrate "consistent" superior accuracy, and enable researchers to examine the development of expertise in face identification over time.

TESTING PRINCIPLE 3. SUPERIOR PERFORMANCE MUST BE ESTABLISHED RELATIVE TO AN APPROPRIATE BENCHMARK

Our definition of expertise specifies that an expert must demonstrate superior performance – but superior performance to what? What is the appropriate benchmark against which expertise should be measured? What should the performance criteria be for someone to be

[43] see Ramon M, Bobak AK & White D (2019) Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, *110*(3), 461-479.

considered an expert? Although these are central questions when applying our definition of expertise in any given context, the workshop members decided not to specify a single benchmark in our definition. This was to enable a generalised definition that was not specific to any one application of face identification.

In most applications, the criteria for expertise would be superior *accuracy* (or fewer errors) relative to the average person on the same task. This benchmark is the accepted standard in psychological studies of expertise, and is also a heuristic used in court, where experts should have expertise that is beyond the average jury member.

But this is not the only potential benchmark. The choice of benchmark may depend on the context in which the claim to expertise is being made. For example, decisions about whether to automate a face identification process may require the accuracy of human experts to be compared to the accuracy of algorithms, or between different algorithms on the market. Similarly, comparison of expertise in different humans might be necessary to make recruitment decisions, or determine which expert's evidence should be weighted more heavily, given the precise parameters of the face identification task.

2.3 KEY SCIENTIFIC FINDINGS

The workshop presentations and discussions identified several key research findings with implications for the assessment and conceptualization of face identification expertise. We summarise these main findings below and provide guidance on the level of scientific support for each.

KEY FINDING 1. WHILE THERE IS CLEAR EVIDENCE OF EXPERTISE IN SOME PRACTITIONER GROUPS, MANY PRACTITIONERS SHOW EQUIVALENT ACCURACY TO AVERAGE PEOPLE

A recent meta-analysis of face identification practitioners shows that while some groups satisfy our definition of expertise (i.e. forensic facial examiners and super-recognisers), others do not (see Figure 4)[44]. For example, tests of passport issuance officers, police officers and border agents have shown accuracy at the same level as novice university students.

[44] see footnote 27.

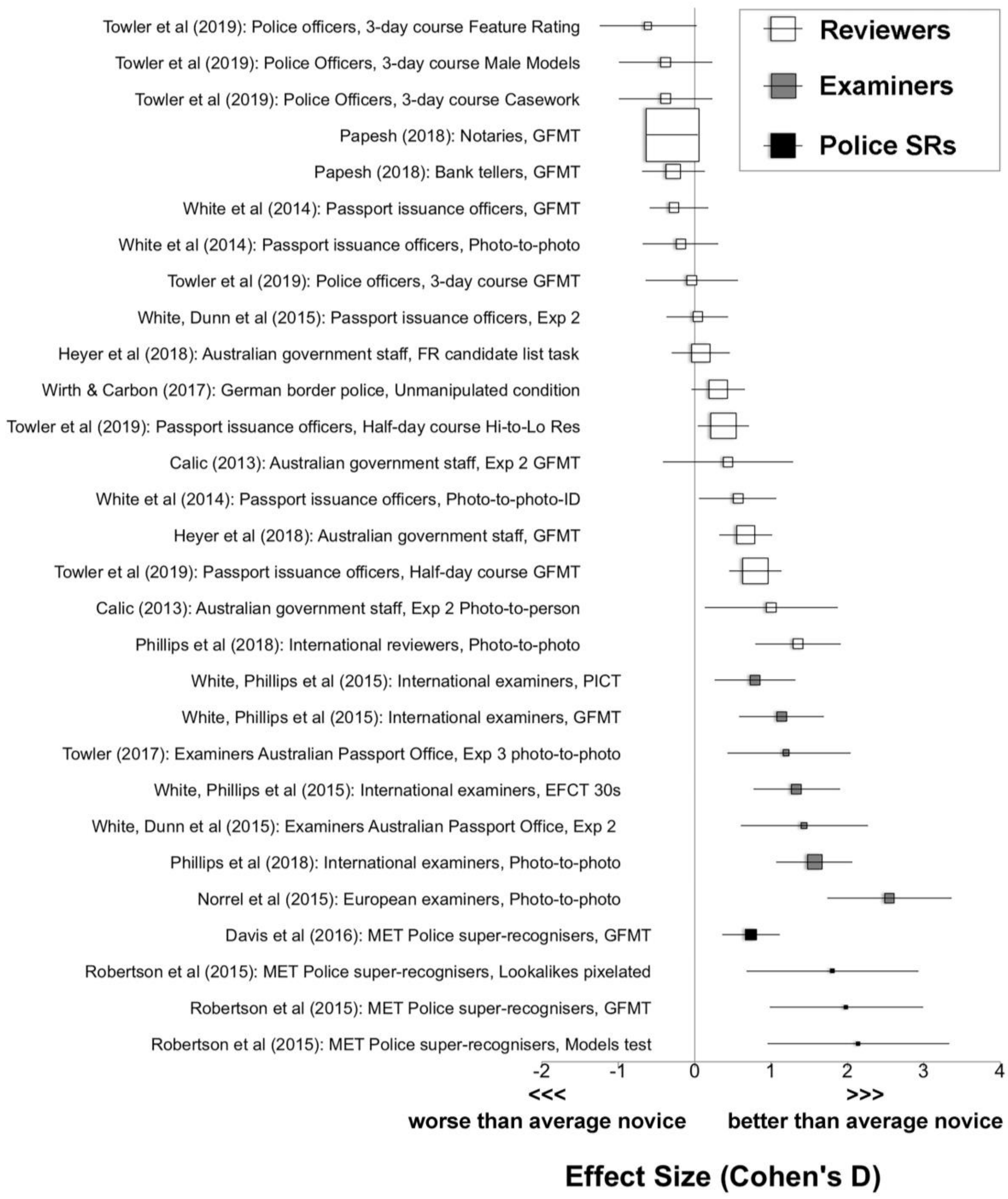

FIGURE 4. This figure shows the size of difference between novices and face identification professionals in 29 unfamiliar face matching tasks taken from 12 peer-reviewed scientific papers. A positive value of 'effect size' (right of the vertical line) indicates that the professional cohort outperformed novices. Where the 95% confidence intervals – i.e. the horizontal lines intersecting the data points – do not overlap with zero (vertical line), this indicates the comparison was not statistically significant. The size of the professional cohort tested is represented by the size of each data point. This figure shows that the vast majority of 'facial review' staff do not perform better than the general population. However, Forensic Examiners and Super-recognisers show reliably superior accuracy to the general population. Details of each of these professional groups are provided in the main text (see Key Finding 2) and further details of these studies are provided in the online publication linked here.

An important implication of these findings is that merely performing face identification tasks on a daily basis, even after receiving training, is insufficient evidence of expertise. All the professional groups listed above received at least basic training, and facial image comparison was a core component of their daily work.

KEY FINDING 2. QUALITATIVELY DISTINCT ROUTES TO EXPERTISE

The existence of the two expert groups – forensic facial examiners and super-recognisers – described above raised important discussions in the workshop related to the conceptualisation of face identification expertise. Both forensic examiners and super-recognisers meet our agreed criteria for expertise. Critically however, forensic examiners and super-recognisers acquire their expertise via different routes[45].

On one hand, forensic facial examiners receive extensive and specialised training in facial feature comparison and deliberate practice of image comparison methods. Examiners are not typically subject to formal selection for these roles via tests of face identification ability. On the other hand, super-recognisers are selected based on pre-existing ability and do not typically receive formal training in face identification. Because these two groups meet our criteria for expertise based on qualitatively different *types* of expertise, this entails that additional guidance is necessary about the relative strengths and weaknesses of these types of expertise.

One distinction that workshop members agreed was potentially useful was between *natural ability* and *acquired ability*. Natural ability was taken to be an intrinsic propensity for high performance that is determined by our genotypes and/or phenotypes. The contribution of this type of expertise cannot be gained through training, study, professional experience, or deliberate practice. Instead, this type of perceptual expertise is gained via our visual experience in daily life, and the necessity to identify faces of people we know, over the normal course of development. Recruitment solutions targeting super-recognisers are aiming to identify people that have developed this natural ability.

Acquired ability is something that can be gained through training, study, professional experience, or deliberate practice and so could be gained in professional settings. It is assumed that acquired expertise is what underlies the expertise of forensic facial examiners. However, future work is necessary to verify the positive benefits of training and professional experience, given that simple before-and-after evaluations of training effectiveness show limited benefits[46]. While some benefits of certain training approaches have been shown to improve accuracy on novices in experimental work, these benefits appear small and would not account for the full accuracy superiority demonstrated by forensic examiners. In general, therefore, improved understanding of natural and acquired ability, and interactions between them, are necessary to improve the assessment and development of expertise in future.

[45] see footnote 16.

[46] see footnote 30; Dolzycka D, Herzmann G, Sommer W & Wilhelm O (2014) Can training enhance face cognition abilities in middle-aged adults? PloS one, 9(3), e90249.

KEY FINDING 3. INDIVIDUAL DIFFERENCES IN NOVICES REFLECTS A STABLE AND HERITABLE COGNITIVE APTITUDE FOR FACE IDENTIFICATION

Relative to most other perceptual and cognitive skills, a person's ability to identify faces is remarkably stable and specific to faces. Reliability of face identification ability compares very favourably to other cognitive skills: it is approximately as high as IQ, which is the most common psychometric measure used to assess individual differences in cognitive ability. Importantly, face identification ability is also a distinct cognitive ability from IQ which means it must be measured separately[47].

A substantial body of scientific research shows that a person's face identification ability is: (i) stable over time, (ii) specific to face identification, in that it does not appear to generalise to other types of objects, and (iii) driven by genetic factors. As a result, face identification ability can be thought of as a 'cognitive trait', largely determined by a person's genes, that can be measured with a high level of reliability.

By implication, recruitment based on reliable psychometric tests offers a very promising strategy to improve the accuracy of face identification in applied settings. However, natural ability in the task is rarely considered when recruiting face identification specialists. Another implication is that large organisations are very likely to have super-recogniser employees in other, non-face identification domains, who could be tested and reassigned to perform tasks suitable to their innate skill.

KEY FINDING 4. THERE ARE LARGE INDIVIDUAL DIFFERENCES IN ACCURACY OF NOVICES AND HUMAN EXPERTS

In *every* group of human participants that have completed face identification tests, researchers have found large, stable variation in individual accuracy scores – some people perform perfectly, while others perform as if they were generating random guesses. Surprisingly, this variation is observed even in the best-performing groups, including professional experts and groups of novices that have been selected based on outstanding ability on face identification tests. Indeed, there is often considerable overlap between the performance of individual novices and people in professional 'expert' groups.

The research indicates that while some professional groups outperform novices, there is no guarantee that *individual group members* will meet our criteria for expertise. In light of this issue, it will be important in many legal and applied settings to test expertise at the level of *individuals*, rather than only at the level of the specific role or discipline[48].

It is not yet clear what causes the high level of variation in accuracy across individual experts. It may be explained by differences in *acquired* ability in individual experts, for

[47] Wilmer JB, Germine L, Chabris CF, Chatterjee G, Williams M, Loken E, . . . Duchaine BC (2010). Human face recognition ability is specific and highly heritable. Proceedings of the National Academy of Sciences, 107(11), 5238-5241.
[48] President's Council of Advisors on Science and Technology (2016) Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods.

example, due to differences in their training, standard procedures or professional experience. Alternatively, or in addition, it may be explained by differences in *natural* ability that existed before beginning professional practice. Discriminating between these accounts is a critical question for future research.

KEY FINDING 5. CURRENT ALGORITHMS SURPASS AVERAGE HUMANS AND ARE COMPARABLE TO THE BEST HUMAN PERFORMERS, BUT ACCURACY VARIES SUBSTANTIALLY FROM ONE ALGORITHM TO THE NEXT

The best facial recognition technology is now as good as the very best human performers, including forensic facial examiners and super-recognisers[49]. Importantly however, in the same way that human accuracy varies substantially from one person to the next (see Key Finding 4), so too does algorithm accuracy. In the most recent evaluations of facial recognition technology, the proportion of errors made by different commercially available algorithms has varied by orders of magnitude. This makes it critically important that end-users of facial recognition technology select the best algorithm for the face identification tasks performed by their organisation.

Selecting appropriate algorithms is particularly important because the performance of algorithms varies depending on the types of face images in question. While one algorithm may perform well on images of female or East Asian faces, another may be better with male and Caucasian faces. This type of differential performance also applies to the type of task parameters in any given organisation. For example, matching passport images requires an algorithm that can tolerate substantial differences in age between images, whereas a security system regulating access to a workplace building would not have this constraint, due to differences in the frequency with which staff cards and passports are issued. Algorithms have also shown differential performance with respect to age differences in images[50].

So, in the same way that a human expert must be tested in a way that directly tests their specific *claim to expertise* (see Testing Principle 1), facial recognition algorithms must also be tested in this way, to ensure that the technology is suited to the particular task that it is performing[51].

KEY FINDING 6. AGGREGATE RESPONSES MADE BY HYBRID HUMAN-AI SYSTEMS PRODUCE OPTIMAL ACCURACY

Face identification systems used in government and industry often involve cascading decisions of algorithms, and different staff members who each judge the identity of faces.

[49] see footnote 14.

[50] Michalski D, Yiu SY & Malec, C (2018) The impact of age and threshold variation on facial recognition algorithm performance using images of children. In 2018 International Conference on Biometrics (ICB), IEEE, pp. 217-224.

[51] Noyes (in press), footnote 6.

While the focus of the workshop was on evaluating the expertise of individual humans and algorithms, there were also important discussions about the possibility of hybrid human-AI systems that involve aggregation of responses from distributed decision-makers (see Hybrid Human-AI Expert Systems).

Recently, researchers have begun to ask how the independent face identity decisions made by humans and algorithms might be aggregated to boost accuracy of end-to-end face identification systems. Aggregation of multiple decisions has significant potential to improve the accuracy of current systems for two reasons. First, due to a statistical phenomenon known as 'wisdom of crowds', averaging decisions from multiple humans boosts the accuracy of face identification decisions[52]. Second, data 'fusion' of match scores produced by multiple algorithms also boost accuracy in many circumstances[53].

In a recent study[54], researchers tested the possibility of 'fusing' judgments of human experts and leading facial recognition technology. By simply averaging the identity judgment decisions made by a leading algorithm and a forensic facial image comparison expert, researchers found that accuracy was increased by 5% relative to either the human or the machine working alone. This promising result suggests that future systems that combine human and computer decisions in intelligent ways can provide optimal levels of accuracy in end-to-end face identification systems.

KEY FINDING 7. HUMANS AND ALGORITHMS PERFORM DIFFERENTLY WHEN MATCHING FACES OF DIFFERENT DEMOGRAPHIC GROUPS

Scientific studies spanning many decades show a person's accuracy on face identification tasks is affected by their own perceptual experience with faces. A prominent example of relevant perceptual experience is the 'other-race effect', whereby people are poorer at identifying races of another ethnicity than their own. This effect is robust, has been replicated many times, and is widely accepted in the scientific community. It has also been shown to affect real-world identification accuracy, where eyewitnesses are more likely to make false identifications when the perpetrator is of another ethnicity to their own.

The leading account of the other-race effect is that it stems from differential levels of contact and perceptual exposure to faces of own versus other ethnicities over the course of development[55]. In turn, the perceptual and cognitive mechanisms involved in face identification end up being tuned to a person's unique perceptual experience. As predicted by this account, differential accuracy has also been shown for faces that are of a different

[52] White D, Burton AM, Kemp RI & Jenkins R (2013) Crowd effects in unfamiliar face matching. Applied Cognitive Psychology, 27(6), 769-777.

[53] see footnote 15.

[54] see footnote 14.

[55] Kelly DJ, Quinn PC, Slater AM, Lee K, Ge L & Pascalis O (2007) The other-race effect develops during infancy. Psychological Science, 18, 1084-1089.

age group to the viewer – younger viewers are poorer at identifying older faces than faces of their own age group, and vice versa.

Analogous effects are also observed in facial recognition algorithms. There, the effect likely arises from differences in the images that algorithms are trained on, and the properties of the test images algorithms are subsequently tuned to identify. For example, in a test of algorithms from the 2006 NIST Face Recognition Vendor Test[56], researchers found that algorithms developed in Asian countries performed comparatively well on Asian faces whereas algorithms developed in Western countries performed better with Caucasian faces. More recent comparisons involving over 100 facial recognition algorithms submitted to the 2019 NIST FRVT show that there are wide discrepancies between how individual algorithms perform across image sets of different races[57].

This recent NIST-led international benchmarking test also showed substantial diversity in the differential accuracy of individual Deep Learning algorithms across different ethnicities, and gender. This test shows that each of the current leading facial recognition algorithms has a slightly different performance profile with respect to faces from different demographic groups. Importantly, more detailed examinations of the causes of these differential effects suggest that it stems from a variety of factors including both the physical characteristics of demographic groups as well as the quality of the representation used by the algorithm[58]. Critically, because of these differences, it is not possible to set a single 'decision threshold' that achieves the same level of accuracy for all demographic groups[59]. Instead, it may be necessary to set different thresholds for different demographic groups (see System Design Consideration 2).

Within the Australian context, it is notable that most studies of demographic bias in face identification – concerning both human and algorithm performance – have examined biases between the largest ethnic groups in Western societies; typically, Caucasian, Asian and African faces. Importantly, there can be substantial variation within these broad groups and there are some ethnicities, notably Indigenous Australians, for which bias in face identification tasks has not been measured.

KEY FINDING 8. ERRORS IN FACE IDENTIFICATION ARE UNAVOIDABLE

A final key finding that the workshop agreed was important is that errors are unavoidable. Even for the best performing expert groups – forensic examiners, super-recognisers and algorithms – it is extremely rare to find individuals who do not make errors in challenging

[56] Phillips PJ, Jiang F, Narvekar A, Ayyad J & O'Toole AJ (2011) An other-race effect for face recognition algorithms. ACM Transactions on Applied Perception (TAP), *8*(2), 1-11.

[57] Grother P, Ngan M & Hanaoka K (2019) Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects, National Institute of Standards and Technology (NIST).

[58] Cavazos JG, Phillips PJ, Castillo CD & O'Toole AJ (2019) Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?. arXiv preprint arXiv:1912.07398.

[59] see footnote 58; see also Krishnapriya KS, Albiero V, Vangara K, King MC and Bowyer KW (2020) Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone. IEEE Transactions on Technology and Society, 1(1), pp. 8-20.

tests. And where those individuals are tested on more than one test, we do not know of *any* single participant that has not made errors.

2.4 CONSIDERATIONS FOR DESIGNING END-TO-END FACE IDENTIFICATION SYSTEMS

While it is critically important to establish expertise by testing performance on face identification tasks, it is also important to test the accuracy of end-to-end face identification systems. Some end-to-end face identification systems will contain only human decision-makers, and others will rely entirely on an algorithm in a “lights out” system. But many will contain multiple human and multiple algorithm decision-makers (i.e. hybrid human-AI expert systems).

Such hybrid human-AI systems are commonplace in applied settings. For example, in an identity management system designed to issue passports to citizens, an algorithm may initially search for the applicant’s face in a database of previously issued passports, before returning a set of potential identity matches that a human adjudicator must review to decide if any match the applicant (see Figure 3B). This human adjudicator may then escalate a decision to a more specialised facial image comparison expert (e.g. facial examiner) where they believe a match may exist.

Other examples of end-to-end face identification systems include forensic laboratories and live facial recognition CCTV systems. In a forensic laboratory, single face identification decisions can be distributed across multiple forensic facial examiners, with standard operating procedures in place to specify how individual observations of feature similarity are aggregated to produce a final identity decision[60]. In a live deployment of facial recognition technology by police, reviewers will monitor a stream of CCTV information and, based on watch list alerts generated by an algorithm, decide whether to stop members of the public and ask them to provide evidence of their identity.

With facial recognition technology’s impressive increase in accuracy over recent years, it is tempting to conclude that humans will soon be obsolete. However, given the powerful benefit provided by fusing decisions of humans and algorithms (see Key Finding 6), optimising the aggregation of human and algorithm decisions in hybrid human-AI systems offers significant potential for further system accuracy boosts.

Moreover, the need for human review may be independent of questions about the relative accuracy of humans and algorithms. Human oversight and guidance of Artificial Intelligence is necessary in many settings due to legal, cultural and legislative requirements, and to ensure the processes are transparent and can be explained to the general public.

[60] Moreton R (in press) Forensic face matching: Procedures and application. In M. Bindemann (ed.), Forensic face matching: Research and practice. Oxford University Press.

Given the importance of considering the overall design of end-to-end face identification systems – over and above the individual accuracy of system components – the workshop members discussed key considerations when designing these systems. The main considerations are as follows.

System Design Consideration 1. PROTOCOLS FOR TESTING THE ACCURACY OF END-TO-END FACE IDENTIFICATION SYSTEMS SHOULD BE ESTABLISHED

The workshop members agreed that academics, algorithm developers and end-users should work together to agree on testing protocols for testing the accuracy of end-to-end face identification systems, i.e. the entire system from start to finish, not just the accuracy of each human/algorithm component contained in the system. The approach to testing outlined in Section 2.2 focuses on the measurement of accuracy in individual experts and computer algorithms. This is the approach that is currently used in most tests of experts in scientific studies, and also in benchmarking tests of computer algorithms. Although this type of testing is critically important, an exclusive focus on individual performance causes a gap in understanding of overall operational performance in systems that incorporate human and algorithm decision-making.

There has been some initial work evaluating the accuracy of hybrid human-AI systems. Simple estimates of the accuracy of such systems can be made by knowing the accuracy of the algorithm and the accuracy of human processing components. In a one-to-many search, algorithms are used to search for a target face amongst a large database of faces, and the algorithm will typically return a list of top matches which are displayed in a 'gallery' to human reviewers for processing (see Figure 3B). While performance of algorithms on this task is published in annual benchmark tests by NIST, there is very little testing of the human review stage of the process, which ultimately determines the accuracy of the decision made by the organisation.

An initial test measuring accuracy of professional staff who perform this review task in daily work as passport issuance officers was published in 2015[61]. These staff made 1 in 2 errors on this task, and specialist facial examiners made 1 in 4 errors. A simple analysis of this result indicates that the accuracy of the system is *doubled* where highly specialist staff are deployed to review the galleries returned by an algorithm, but this is often not the case in applied settings.

The test of passport issuance officers by White and colleagues (2015) described above was conducted using a simulation of their normal face identification workflow, but was conducted 'offline', outside of normal working conditions where staff knew they were being tested. In future, it may be necessary to employ both 'operational' and 'offline' testing protocols. Given the complexity of many operational tasks, it can be challenging to measure the accuracy of

[61] see White et al.(2015), footnote 6.

systems in normal operational conditions. However, this approach provides additional benefit because staff are unaware they are being tested[62].

Establishing protocols for operational testing may not be feasible in many settings (see also Considerations for designing end-to-end face identification systems). Another option is therefore to develop 'test systems' that mirror the operational conditions, but are not part of the operational workflow, and where system administrators have control of the image data contained in the test environment. These test systems would function like the normal system, using image data sampled from normal operation, but would enable precise cognitive tests to be carried out.

Where these types of test systems can be integrated into the normal workflow of operators, it may also be possible to carry out covert tests of performance that enable regular system testing. Such a strategy is challenging to implement, but entirely feasible – it is regularly used in airport baggage screening to maintain acceptable levels of accuracy, and is starting to be implemented in the forensic sciences.

As with the broader testing principles outlined in Section 2.2, input from behavioural and cognitive scientists will be important for establishing system testing protocols. However, given the complexity of these systems, it is necessary for interdisciplinary teams of scientists, practitioners, IT system designers, analysts and software developers to work together to provide these solutions.

SYSTEM DESIGN CONSIDERATION 2. ALGORITHM THRESHOLD SETTING IS A CRITICAL POLICY DECISION WITH IMPLICATIONS FOR THE EFFECTIVENESS, COST-BENEFIT TRADEOFF, AND FAIRNESS OF FACE IDENTIFICATION SYSTEMS

Facial recognition technology works by comparing two images and returning match scores that indicate the degree of visual similarity between them. A decision threshold must then be applied that determines the degree of similarity required for a face to 'match'. In the case of identity 1-to-1 verification, for example used in automated border gates (see Figure 3A), the decision threshold dictates whether the automated gates admit the person (where the match score exceeds threshold) or diverts them to a primary line officer (where the match score is less than threshold). In the case of 1-to-n search systems, the threshold setting dictates the faces that are presented for human review in a candidate list (see Figure 3B). This type of system is typically used for identity fraud protection, police investigation and live 'watchlist' surveillance.

Decision thresholds are set manually by the system administrator. This a critical consideration because it affects the *effectiveness, cost-benefit tradeoff, and fairness* of face identification systems, in the following ways:

[62] For example, see Addressing Significant Vulnerabilities in the Department of State's Passport Issuance Process. (2009) US Government Accountability Office.

- ***Effectiveness.*** To ensure optimal accuracy, thresholds should be set based on an analysis of the distribution of match scores for a given system in operation (see Figure 3). While vendors typically provide some guidance on appropriate threshold settings, the administration of this falls to end-users. And beyond a basic analysis of match scores on a sample of operational image data, there are likely many fine-grained considerations that have a substantial impact on accuracy and should be tested by end-users.

  For example, the number of years that have passed between two images of the same face in a dataset will have a drastic effect on the distribution of match scores for matching faces. As mentioned previously, the number of years between images will vary from one task to another. For example, passports tend to have longer validity than workplace identification documents. Where there are large differences in the age of images, this will cause matching faces to produce more variable match scores, and so it is likely that thresholds will need to be set differently in these example scenarios.

  Match distributions may also vary systematically within a given system depending on the type of identification decision. For example, the validity of passports can depend on the type of application (e.g. first passport vs. renewal), and so age differences expected in these application types may be different. Optimal accuracy in this system would require different thresholds for different application types. Similarly, match distributions may vary by age or ethnicity of the applicant, and in these cases, optimal systems will vary thresholds as a function of the applicant demographics (see also 'Fairness' below).

- ***Cost-benefit tradeoff.*** Setting an appropriate threshold to produce optimal algorithm accuracy is not a simple task. For example, in a one-to-many database search application, the threshold selected will determine the amount of human processing required (see Figure 3B). Where a high threshold is set, relatively few faces will be selected as potential matches for humans to review, potentially missing some genuine matches. Where a low threshold is set, many more cases will require human review, thereby increasing staff costs but reducing the risk of missing genuine matches. The result is an econometric trade-off between cost and risk whereby the cost of extra scrutiny of matches offsets the risk of, for example, arresting an innocent person. Despite improvements in accuracy and understanding of human and algorithm performance, errors continue to be inevitable and so risk mitigation is a critical consideration in any face identification system.

- ***Fairness.*** Face identification decisions can have negative impacts on people's lives. For example, by affecting their access to government services, or the level of interrogation they receive at national borders, and can even lead to wrongful arrest. It is important that face identification systems are designed so they do not have unnecessary negative impacts for people, and that these impacts are not disproportionately felt by certain sectors of society or demographic groups.

Threshold setting constrains the fairness of face identification systems in many ways. As described above, a high threshold setting will increase the likelihood that a genuine match will be missed by a system (see Figure 3). Conversely, a low threshold will increase the likelihood that a non-matching individual will be falsely matched to another face, which could lead to police questioning, travel disruption or more serious outcomes for the person in question.

To ensure equitable and fair systems it is important to consider setting separate thresholds for different demographic groups. Recent work has shown that match score distributions vary substantially between different ethnic groups, genders and age groups (see also Testing Principle 1)[63]. As a result, setting a single threshold for all these groups will lead to differential accuracy for each subgroup, which can potentially lead to bias in important legal decisions and in interactions between citizens and governments.

SYSTEM DESIGN CONSIDERATION 3. FUTURE FACE IDENTIFICATION SYSTEMS CAN BE BUILT TO IMPROVE HUMAN-COMPUTER INTERACTIVE PROCESSING

Scientific research shows that the most accurate face identification decisions are made by combining match scores generated by humans and algorithms. By simply averaging the independent judgment of humans and algorithms, near-perfect accuracy was achieved on a recent challenging test[64]. This accuracy was higher than was achieved by either leading forensic facial examiners or leading facial recognition technology working alone.

Current face identification systems do not typically operate by aggregating the responses of humans and algorithms. Instead, each of these components make serial decisions, with algorithms typically used to pre-screen face identification decisions. Humans then process the exception cases where algorithm certainty in the decision is low or adjudicate a candidate list of images that have surpassed the match score threshold (see Figure 3).

In future, designers of face identification systems should consider ways to implement a simple averaging of independent human and algorithm decisions. For example, in a forensic context, facial examiners may often receive images that have been matched by face recognition technology at some prior stage. As far as we are aware however, they do not aggregate their own independent judgments of image similarity with the algorithm's match scores, despite evidence this would benefit accuracy.

This type of response aggregation may also be feasible in candidate list review tasks (see Figure 3B). For example in the context of police investigations, when police officers use facial recognition technology to search mugshot databases using photo evidence, they are often presented with a large array of potential matches and are required to decide if any of

[63] see footnote 57.
[64] see footnote 14.

those faces ‘match’ a suspect. Alternatively, a system could be designed that presents a smaller list of the top matches in a random order and requires a human reviewer to rate the similarity of each to the suspect in sequence, before fusing the human judgment/s with algorithm match scores. This approach would limit the likelihood of making false matches that waste police resources in subsequent investigation of these leads, adding more evidentiary value to the leads that are generated by this process.

Designing information processing workflows that enable aggregation of independent identity judgments made by humans and algorithms promises to boost the operational accuracy of face identification systems – perhaps more than is currently possible through training and recruitment[65]. However future research on this topic is recommended because the benefits of this approach have not been tested in operational tasks.

SYSTEM DESIGN CONSIDERATION 4. THE RELATIVE FREQUENCY OF FACE MATCHES/NON-MATCHES AND THE BROADER DECISION-MAKING CONTEXT

Human face identification accuracy depends on the broader context in which decisions are made. For example, image comparison decisions may be sent to an expert forensic examiner because they have been referred by other staff or flagged as potential matches by facial recognition software. The origin of the comparison affects the prior likelihood that the images are indeed a match, irrespective of the content of the images themselves.

Similarly, when 1-to-many face identification systems are used in criminal investigation (Figure 3B), the image database searched by the algorithm will have a substantial effect on the base-rate probability of genuine matches. For example, highly similar face returned from a state-database search is more likely to be the target than the same face returned from a national-database search, because the latter would include many more non-matching faces that look extremely similar to the target.

For 1-to-1 verification systems used at border control (Figure 3A), the likelihood that a traveller – who has been referred to a border control officer by automated border control gates – has a false passport is greater if the referral is due to a low match score as opposed to referral because of a failure of the border gate to read the passport chip.

These are important considerations because fluctuations in the ‘base-rate’ probabilities of match and non-match pairs being encountered – or perceived fluctuations caused by the situational context – affect human decisions. Variations in the prevalence of match pairs in a series of face matching decisions affects accuracy because people are more likely to miss rare events than common ones.

[65] Balsdon T, Summersby S, Kemp RI, White D (2018) Improving face identification with specialist teams. Cognitive Research: Principles and Implications, 3(25).

There are ways to avoid missing rare events, such as travellers who are using a false identity document. For example, a successful strategy to mitigate these errors in x-ray security screening at airports is by projecting synthetic prohibited items – for example, a knife – onto the operator's visual display via computer software[66]. Similar systems may help to mitigate face identification errors in security tasks where staff make high volumes of face matching decisions. In addition, because the context surrounding decisions can influence the perceived likelihood of encountering a match/non-matching image pair, systems that exert control over the context that operators are exposed to can help mitigate the negative effects of context. For example, in forensic science more broadly, 'sequential unmasking' techniques are used to limit exposure to other case facts that are likely to influence a forensic examiners' image comparison judgments[67].

There is limited research on the effects of these types of contextual influences on face identification decisions [68] . Indeed, the science of face identification typically tests performance in conditions that strip away any sources of contextual influence. This creates a mismatch between scientific knowledge of performance and the rich contexts in which real-world face identification decisions are made. This is an important area for future research.

System Design Consideration 5. A NEW TYPE OF EXPERT – A FACE IDENTIFICATION SYSTEM ANALYST – IS REQUIRED TO HAVE BROAD UNDERSTANDING OF HUMAN AND ALGORITHM PERFORMANCE

The workshop was focused on defining and evaluating expertise in performing face identification tasks. It remains critically important to have experts in face identification performing key roles, whether they be forensic scientists, passport issuance officers or border control agents. But, there is also a need for experts that oversee the entire end-to-end face identification systems these individual experts work within, and that can explain the operation of these complex systems. We term this role: *face identification system analyst*.

Key components of the *face identification system analyst's* role include: (i) designing procedures to test system performance using image data and conditions that are representative of normal operations; (ii) analysing performance data from human and algorithm decision-making; (iii) adjusting system design based on this analysis to optimize system performance; (iv) explaining how the overall system and its components arrive at face identification decisions, and (v) communicating with stake-holders and decision-makers about the design, validity and reliability of the system. This means the role of the face

[66] Cutler, V., & Paddock, S. (2009). Use of threat image projection (TIP) to enhance security performance. In Security technology, 2009. 43rd Annual 2009 International Carnahan Conference (pp. 46-51).

[67] Krane DE, Ford S, Gilder J, Inman K, Jamieson A, Koppl R, Kornfield I, Risinger DM, Rudin N, Taylor MS & Thompson WC (2008) Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation, Journal of Forensic Sciences, 53(4), 1006–1007.

[68] but see Papesh MH, Heisick LL & Warner KA (2018) The persistent low-prevalence effect in unfamiliar face-matching: The roles of feedback and criterion shifting. *Journal of Experimental Psychology: Applied*, *24*(3), 416; Fysh MC & Bindemann M (2018) Human-Computer Interaction in Face Matching. *Cognitive Science*, 42(5), 1714-1732.

identification system analyst requires a high level of technical competence, combined with an understanding of behavioural research methods, including test design and statistical analysis training. Alternatively, this role could be filled by a generalist who manages an interdisciplinary team of experts in each of these areas.

PART 3: THE FUTURE OF FACE IDENTIFICATION: ADVANCING AN INTERDISCIPLINARY FIELD BY SYNERGIZING RESEARCH AND APPLICATION

The workshop was designed to reconcile strands of face identification research and operational practice that have historically operated independently and with very little collaboration. Our primary aim was to provide a unifying definition of expertise in face identification, but more broadly, we hoped to find a common purpose based on shared principles that can advance theory and practice in this area in the years ahead.

During the workshop, a picture of a new interdisciplinary field of research emerged – one that intersects psychology, computer science, forensic science and law. The field of *face identification* is emergent from the practical problems associated with identifying unfamiliar faces in the modern world. However, this does not mean it is strictly an applied field of study. Academic research in face identification is often separated along traditional lines of theoretical and applied work, but it was clear from conversations between applied and theoretical researchers that the success of both research trajectories depends on one another.

Key to the development of this field then will be collaboration. Collaboration between academics working in the intersecting fields. Collaboration also between these academics and the practitioners and policy-makers who are tasked with implementing accurate, cost-effective, and fair face identification systems. Critically, the development of this field relies on a two-way conversation in which academics provide recommendations, but critically requires academics to tailor their research questions based on feedback from practitioners and policy-makers in ways that make the work they do applicable and useful in applied settings.

Perhaps more important than collaboration, interdisciplinary training and career pathways for the next generation of researchers, practitioners and system analysts are necessary to meet future needs. As should be clear after reading Parts 1 and 2, future researchers and practitioners in this field will need to be conversant in multiple discipline areas.

In Part 3 we reflect on discussions from the workshop that might help to guide this emerging field in the years ahead. First, we outline the ways in which collaboration can continue to help meet the aims of practitioners and policy-makers. Second, we consider the questions that researchers can address now that are likely to lead to improvements in face identification practice. Third, we consider how post-workshop activities might be designed to sustain collaboration and promote development of future leaders in this field in the medium- to long-term.

3.1 WHY RESEARCH COLLABORATION IMPROVES PRACTICE

IMPROVING FORENSIC SCIENCE

Following high-profile errors, forensic science has been under increased scrutiny, most prominently by expert panels assembled by the US Government[69]. These panels aimed to assess the validity and reliability of forensic feature comparison disciplines (e.g. fingerprint comparison, bite-mark analysis). They concluded that with the exception of fingerprint comparison and single-sample DNA analysis, these disciplines do not have the necessary scientific backing to support the claim that they can reliably identify individuals. They recommended forensic scientists work with cognitive and behavioural scientists to address the lack of empirical evidence supporting the validity of these approaches.

Face identification was not included in these reports, but the workshop agreed with their conclusion that collaboration between forensic science researchers, practitioners and the broader scientific community is essential to progress. The workshop also agreed that face identification has a substantially stronger footing on which to claim it is valid and reliable than many of the disciplines assessed by the US expert panels. This is because of published empirical proficiency testing of practitioners, reliable automated methods and international best-practice guidelines. In addition, initial contrasts between novices and forensic facial identification experts in scientific publications point to high accuracy in some groups of practitioners.

Importantly, much of this recent empirical work has been driven by cognitive, behavioural and computer scientists rather than solely from within the forensic science community. As a result, many of the key findings presented in this digested analysis have not yet been incorporated into practitioner guidelines for forensic science, and much of the knowledge generated by researchers is not widely known to practitioners. So in the years ahead, greater collaboration between forensic scientists and researchers should be a priority to ensure that evidence-based procedures are adopted in face identification practice.

Critically, this translational work – from research to practice – must be part of a broader cycle of knowledge development that incorporates reciprocal knowledge exchange. Feedback from practitioners to researchers is also fundamental to this process, partly because without a detailed picture of working practices of forensic examiners, researchers are unable to develop appropriate tests to validate expertise[40].

For example, facial forensic examiners routinely perform a number of sub-tasks for each image comparison task. They will typically first assess the quality of the imagery in question and whether it is suitable for an identification decision. They will then typically complete

[69] President's Council of Advisors on Science and Technology (2016) Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods; see also National Research Council (2009) Strengthening forensic science in the United States: A path forward

detailed comparison of features, noting similarities and differences. Only then will they make a final judgment of the degree of support in the image evidence for the images being of the same person. Because each of these sub-tasks may rely on separate skills, it is important that these processes are subject to empirical performance testing. But without detailed knowledge of working practice, scientists are unable to create valid tests (see Testing Principle 1 for further discussion).

This cycle of translational (research ⇨ practice) and reverse-translational work (practice ⇨ research) has been key to the early progress made in establishing the empirical basis for the forensic science of face identification (e.g. see Key Finding 1 & Key Finding 2). It is critical that this cycle continues to ensure that researchers can create valid tests that evaluate claims to expertise directly (Testing Principle 1). When research identifies that forensic practitioners are unable to perform a given task under given conditions, then this provides evidence that this practice should cease – at least until research identifies ways to reliably improve accuracy. Practitioners need to know what they can and cannot make claims about, what forensic practices are okay to use as evidence and which are not. Critically, empirical research needs to inform this process.

In addition to working with forensic scientists, there is also a need for scientists, practitioners and policy-makers to work together to optimise the way face identification evidence is presented to judges and jurors in court. At the moment there is a disconnect between face identification evidence that appears in courts and the science of face identification. And there is also a gap between work that skilled face identification practitioners in government and police do and the type of expert evidence that appears in court, which is often provided by independent experts that do not adhere to best-practice standards. These gaps point to the need for a more unified field and so closer collaboration is key.

DESIGNING BETTER FACE IDENTIFICATION SYSTEMS

Face identification systems incorporating algorithm and human decision-making are now commonplace in government, security and forensic settings. Simplified examples of these systems are shown in Figure 3. But as should be clear from this report, the reality of operational systems is that they are complex, and optimal functioning depends on the calibration of many system components – algorithm functioning, business decision thresholds, human workflow and decision-making.

The complexity of these systems requires support from specialists with training in statistics, computer science, IT systems, and human behaviour to ensure their effective and accurate operation (see System Design Consideration 5). To create systems that are fair, these systems should also be designed in consultation with legal, human rights and policy experts, and so it is important to establish a common language between these experts and researchers/practitioners with expertise in technical aspects of these systems. Interdisciplinary collaboration is therefore essential to create face identification systems that are accurate, effective, and fair (see System Design Consideration 1 & System Design Consideration 2).

The need for cooperation when studying face identification systems is highlighted in two recent studies commissioned by the London Metropolitan Police Service to measure the operational accuracy of a Live Facial Recognition system[70]. Although these studies used similar methodology to collect performance data, they measure error-rate differently. While the MET Police chose to report error-rate as a function of the total number of people that were scanned by the software, giving 0.1% false false-positive rate, the error-rate reported by the University of Essex was a function of the total number of faces that were flagged by the FR software resulting in a vastly higher 80% false-positive rate. This process highlights the need for practitioners to work together with technical specialists and academics to agree on appropriate measures of accuracy when systems are tested in operation.

Critically, neither method of calculating error-rates described above included false-negatives, because researchers did not have control over the faces that were presented to the CCTV cameras, nor the faces contained in police watchlists. Indeed, the only way for researchers to verify the veracity of correct matches (or 'watchlist alert'), was to stop members of the public and ask for their identity documents. This highlights the complex challenges that are often faced when testing systems in operation, as well as the legal and privacy implications of false matches from facial recognition technology.

An important limitation of these studies is that they aimed to measure *algorithm* accuracy from their observations of operational deployment, as opposed to the accuracy of the *system*. For instance, 16 of the 42 matches in the evaluation by the University of Essex were deemed 'non-credible' matches by police officers monitoring the technology and so were recorded as errors of the technology. It is not clear whether these 'non-credible' judgments were made on the basis of perceptual comparison of the watchlist images to the person on CCTV, or whether some other information was used, such as whether the person was currently imprisoned. This consideration is critical, because watchlist alerts that are deemed not to be matches based on perceptual analysis by the human operator would be considered as correct decisions at a 'system-level', where the system is conceptualised as a combination of algorithm and human decision-making.

Holistic evaluations of face identification systems require interdisciplinary teams of psychologists, computer scientists and others to work together. A priority in this work should be to arrive at common protocols for testing the accuracy of systems. The lack of agreed protocols for calculating end-to-end system accuracy means that this critical information has so far been omitted from reports of operational system tests. Lack of protocol forms a substantial barrier to progress in research and development of face identification systems;

[70] Fussey F, Murray D (2019) Independent report on the London Metropolitan Police Service's trial of live facial recognition technology, University of Essex and National Physical Laboratory (2020) Metropolitan Police Service live facial recognition trials.

in providing appropriate information on which to base policy decisions; and when informing the general public on how facial recognition technology will impact their lives.

Although collaboration is critical, it is unlikely to be sufficient to achieve the level of interdisciplinarity that is necessary to solve these complex problems. To sustain a connection between practice and research in this field, it will also be necessary to train a new generation of researchers and practitioners that are conversant with a broad range of traditional disciplines that intersect this emerging field.

3.2 FUTURE USE-INSPIRED RESEARCH DIRECTIONS

There has been a substantial amount of progress in understanding expertise in face identification over the past few decades. Multiple important future research directions were discussed at the workshop to ensure that this growth in knowledge continues to provide benefits to theory and practice in this field. Below we list some that were discussed in detail. We briefly review work that has already been carried out to address these gaps and what needs to be done in the years ahead.

FUTURE RESEARCH DIRECTION 1. IN COLLABORATION WITH PRACTITIONERS, SCIENTISTS SHOULD CONTINUE TO DEVELOP VALID, RELIABLE AND CALIBRATED TESTS

Substantial progress has been made in recent years creating scientifically validated tests of face identification ability. These tests have been developed to measure ability on various identification tasks: recognising familiar faces, memorising unfamiliar faces, and perceptually matching unfamiliar faces. These tests are reliable measures of face identification ability. A reliable test means that a person who scores highly on these tests is likely to score highly on the same test taken a second time. However, discussions during the workshop raised some issues that may limit their applicability and necessitate further test development in future, for example:

- ***The best combination of tests to assess expertise in applied settings is not clear.*** Reliable measurement tools are available, and performance shows a moderate level of correlation between these tests, suggesting that they tap a general skill in face identification. But a key question is whether they are suited to selecting for the applied tasks. There is currently very limited data on whether people selected using these standard tests *actually* go on to perform well on the real-world task they have been selected to perform. Some tests are likely to be more suited to certain applied tasks. For example, recognising known suspects in CCTV streams is likely to involve a slightly different set of skills to performing detailed comparison of face images. Further, there may be different sets of skills that people in these roles require. A battery of tests is therefore needed to assess an individual's suitability for a given role, but there is currently no agreement on what should be included in that battery, or what the ideal performance profiles on this battery would be for a given role.

- ***Researchers do not have a detailed understanding of the tasks performed in applied settings**.* A major limitation to developing valid tests of professional tasks, and recruitment protocols for professional roles, is that researchers simply do not have a sufficient understanding of the tasks performed in applied settings[71]. Because scientists rarely experience the type of tasks performed in applied settings, there is often a disjuncture between tasks they use in research and those that are performed in professional settings. This disjuncture is especially problematic because there are a diverse set of operational roles that face identification professionals currently perform, and that they will be expected to perform in the future. In collaboration with practitioners, behavioural scientists should therefore aim to analyse the various tasks that are performed in professional settings using 'task analysis' approaches. A task analysis approach will require an increased level of collaboration between academics and practitioners.

- ***Existing tests do not enable repeated testing to examine the development of expertise over time**.* A key outstanding question is whether high levels of accuracy in face identification tasks can be acquired professionally. Whereas simple before-and-after studies of short training courses show that these do not confer immediate improvement, longer courses do show some improvement[72]. However, the higher performance of forensic facial examiners suggests that more extended on-the-job training and mentorship can promote high levels of accuracy. To understand the contributions of training, mentorship and professional experience it is critical to conduct longitudinal testing where the accuracy of professional groups is tracked continuously over their professional career. However, no current tests are available that contain multiple sub-tests of equal difficulty that would enable this type of repeated testing.

- ***Existing tests are not challenging enough to discriminate between the very highest levels of ability.*** A key barrier to improving the accuracy of experts in face identification is to create challenging tests that distinguish between experts with the highest levels of ability. Recent tests have mined datasets of images to find the most challenging pairs using a combination of human and machine performance data. Researchers found that the highest performing human experts and leading algorithms both perform very close to perfect accuracy on these tests[73]. This result suggests that future tests will need to mine larger sets of images, and perhaps images captured in more variable imaging conditions, to make tests challenging enough to discriminate between the highest levels of human and machine accuracy. Preliminary work presented by Jonathon Phillips and Jeremy Wilmer at the workshop has begun to apply Item Response Theory to stratify the difficulty of test items, and further work in this area

[71] Moreton R, Pike G & Havard C (2019) A task-and role-based perspective on super-recognizers: Commentary on 'Super-recognizers: From the lab to the world and back again'. British Journal of Psychology, 110(3), 486-488.

[72] see footnote 30.

[73] see footnote 14.

will be necessary to sustain the requirement for challenging face identification tests in the future.

Future Research Direction 2. MEASURING PERFORMANCE OF ENTIRE END-TO-END FACE IDENTIFICATION SYSTEMS, NOT JUST COMPONENTS, TO INFORM DEVELOPMENT OF BETTER SYSTEMS

An important recent advance has been to consider the combined accuracy of human and machine processing in face identification. Some tests have evaluated the effectiveness of human review of 'candidate lists' returned by facial recognition algorithms[74] (see Figure 3B). Others have examined novel methods of aggregating human and algorithms via 'fusion'[75], showing promising benefits of this type of combination (see Future Research Direction 3).

It is important that future research continues this emergent focus on *system* performance, and there are a number of promising directions for this work, exploring the costs and benefits of different configurations of hybrid human-AI systems. Existing tests of combined human-AI accuracy did not consider the impact of varying algorithm threshold, nor the volume of human adjudication that would be required at each threshold. Neither did they consider the cascading decisions made by different groups of staff. It is very common for face identification staff to refer suspicious or challenging cases onto a specialist team for more detailed analysis. To measure accuracy of face identification systems it is therefore necessary to examine the system as a whole, not only as a collection of isolated components.

Measuring the accuracy of the whole system will enable system developers to design more accurate, efficient, and fair systems. It will also provide important feedback to computer scientists on the operational reality of face identification systems. In the computer science literature, the focus is on algorithm accuracy in ideal conditions. Accuracy of the algorithm in realistic operational settings is rarely (or poorly) assessed, and the impact of human adjudication is never quantified. This disjuncture between the accuracy in academic studies and benchmark tests sets unrealistic expectations of the effectiveness of these systems when deployed in real-world tasks.

Future collaborative work between academics and system designers should therefore aim to assess the accuracy of the entire system, from beginning to end. However, it was noted in discussions that there is often a lack of expertise in organisations that deploy face identification systems to conduct the appropriate testing of algorithms and system performance. Testing operational performance of face identification systems is critical given variations in algorithm and human accuracy. To enable sustainable system-level testing it

[74] see White et al.(2015), footnote 6.
[75] see footnote 15; see also footnote 14.

may be necessary for organisations to seek assistance from biometric testing specialists, or to begin to employ or train this type of expert within their organisation.

FUTURE RESEARCH DIRECTION 3. COMBINING FACE IDENTIFICATION DECISIONS MADE BY PEOPLE AND ALGORITHMS

Initial reports of fusing human and algorithm face identification decisions have shown promising results, with these hybrid human-AI systems achieving accuracy that surpasses either human experts or algorithms acting alone[76]. As we outlined in System Design Consideration 3, future face identification systems can be designed to leverage this accuracy benefit. Similar fusion effects are seen when combining independent judgments made by humans and so there is substantial promise in designing systems that aggregate independent human judgments.

To support this effort, researchers can examine the fundamental basis of these fusion effects in greater detail. The currently accepted cause of fusion effects is that different decision-makers – i.e. different individuals, or different algorithms – approach the task of identifying faces in a different way. For example, identification decisions may be based on different features, or by using different strategic processes to arrive at decisions. Divergence in cognitive strategy causes errors of these different decision-makers to become uncorrelated, and so fusion of their responses serves to 'wash out' these errors by statistical aggregation.

Currently, theoretical understanding of the causes of diversity in cognitive strategies and perceptual representations used by decision-makers is underdeveloped. Work that explores why different people and different algorithms arrive at different representations and strategies for performing face identification tasks may help strengthen the benefits of fusion in future.

FUTURE RESEARCH DIRECTION 4. MEASURING BIAS IN FACE IDENTIFICATION

As discussed in Key Finding 7, humans and machines are 'biased' because they make more or less errors depending on the demographic group. Investigating biases is an important topic for future research to ensure fairness in the treatment of different demographic groups by face identification systems.

The workshop discussed the technical reasons for why humans and face identification systems show bias. These technical discussions hinged on the match score distributions and decision threshold settings that underpin face identification decisions (see Figure 3 for a visualisation of these). From a technical perspective, there are three main reasons why demographic biases may occur.

[76] see footnote 14.

First, it might be that the underlying distributions of match scores differ between demographic groups. In a recent test of facial recognition algorithms, researchers found that match score distributions for different demographic groups differed such that different thresholds were required to produce optimal accuracy in each group (see Key Finding 7). This study suggests that 'bias' of algorithms can be removed by setting appropriate thresholds separately for each demographic group.

Second, the bias could occur because underlying match and non-match distributions overlap more for one demographic group than for another. This is the case for humans, where poorer accuracy with faces of a different ethnicity than our own is caused by a reduced ability to perceptually discriminate between other ethnicity faces. In cases where algorithms also show differences in overlap on match/non-match distributions, then further training and tuning of the algorithm can potentially produce equal performance.

A third potential reason for bias is that the threshold settings vary by demographic group, but the underlying distributions are relatively stable. This type of bias appears to be underlying other types of cognitive bias that have been reported in forensic science domains, for example in fingerprint comparison. Contextual information about a case can bias fingerprint comparison experts to be more likely to make 'non-match' decisions. This is independent of a person's ability to perceptually discriminate between fingerprints, and is instead a 'bias' in the purest sense of the word. Similar biases have been found in face identification due to contextual biases from displays of algorithm match scores, the presence of other biographical information and the relative frequency of encountering match and non-match decisions.

Future research that aims to understand the relative influences of the mechanisms described above in producing differential accuracy in humans, facial recognition technology and hybrid systems will help to inform the design of future systems that do not show differential accuracy for demographic groups.

Future Research Direction 5. EFFECTIVE COMMUNICATION OF UNCERTAINTY IN FACE IDENTIFICATION DECISIONS

Use of response scales to appropriately communicate uncertainty appears to be a feature of expertise. For example, forensic examiners are less likely to make high confidence errors than other groups[77]. Indeed, this kind of conservatism is one of the hallmarks of expertise in the forensic sciences more broadly. But we do not know whether this feature of expert performance is simply a bias towards being conservative, perhaps caused by the fact that forensic experts do not want to make costly errors, or whether it is the result of better calibration of certainty judgments given the strength of evidence. Future research is necessary to discriminate between these alternatives.

[77] see footnote 14.

Another important direction for research is to examine how uncertainty in face identification is communicated, and the effectiveness of this communication in modulating the interpretation of evidence in court and in applied settings. For example, how should experts communicate doubt in their decisions without invalidating the decision or causing people to disregard it entirely? How effective is current expert training in ensuring effective communication of evidence? How and *who* should communicate uncertainty in algorithm match score evidence in court? These are all questions that are deserving of further attention.

3.3 A STEERING COMMITTEE TO SUSTAIN POST-WORKSHOP ACTIVITY

Recent collaborations between computer scientists, cognitive psychologists, forensic scientists and legal experts have helped establish an interdisciplinary face identification field that is characterised by an integrated view of applied and theoretical questions. It is essential that this continues in the future to address the mounting challenges ahead, as face identification becomes increasingly prevalent in forensic, legal and identity management processes. Training and developing a new generation of scientists and practitioners that are 'multilingual' in the discipline areas that intersect this emerging field will also be key to meeting these challenges.

To facilitate this, the workshop conveners have proposed that we establish an international steering committee to ensure that the outcomes of the workshop are communicated broadly and to the appropriate stakeholders, and to promote the broadest implementation of our recommendations possible. Initially, this steering committee would consist of the conveners and workshop leaders. Beyond that, we would invite other stakeholders to join this committee, primarily policy-makers and practitioners from police, government and law. The Steering Committee would meet 2-3 times per year to measure progress against, and develop initiatives in service of, the following aims:

1. Coordinate the dissemination of workshop outcomes to scientists, practitioners, policy-makers, and the general public.

2. Encourage collaboration between scientists, policy-makers and practitioners in face identification.

3. Promote career development of early career researchers and practitioners such that the next generation of leaders are multilingual in the discipline areas that intersect the field of face identification.

4. Facilitate development of the scientific tools and methodologies to support accreditation of professional face identification experts.

5. Disseminate knowledge to inform development of evidence-based professional standards for face identification experts.

6. Explore funding opportunities that can facilitate the aims of the steering committee.

WORKSHOP MEMBERS

CONVENORS

Dr David White

UNSW Sydney, Australia | david.white@unsw.edu.au

David White is a cognitive psychologist studying the perceptual and cognitive processes involved in person perception. Recent work focusses on individual differences in people's ability to identify faces, both in novices and in expert groups such as passport officers, police officers and forensic examiners. He has worked with a range of partners in Australian Government (DFAT, DTA, DST, RBA), police (NSW, MET Police Forces) and international research institutions (NIST) to address problems of applied and theoretical significance.

A/Prof Romina Palermo

University of Western Australia, Australia | romina.palermo@uwa.edu.au

Romina Palermo is interested in understanding the perceptual, cognitive and neural basis of person perception. Research on face identification has focussed on understanding why children and adults vary in their natural ability to recognise face identity, and why some children and adults (e.g., those with prosopagnosia or autism) find it very difficult to identify faces.

Dr Linda Jeffery

University of Western Australia, Australia | linda.jeffery@uwa.edu.au

Linda Jeffery is a cognitive psychologist who studies the visual processes that support face identification with a particular interest in how these skills develop during childhood, understanding why expertise varies considerably among individuals and how face skills may develop differently in those with developmental disorders (e.g., autism) and clinical conditions (e.g., social anxiety).

Dr Alice Towler

UNSW Sydney, Australia | a.towler@unsw.edu.au

Alice Towler is a cognitive psychologist whose research focuses on improving the accuracy and efficiency of face identification systems. She has worked closely with the Australian Passport Office and Metropolitan Police Service (UK) to evaluate the effectiveness of professional training courses and develop new evidence-based training for facial image comparison. She also has a broader interest in improving the evidence-base in the forensic sciences through her work with the Evidence-Based Forensics Initiative.

Prof Richard Kemp

UNSW Sydney, Australia | richard.kemp@unsw.edu.au
Richard Kemp is a forensic psychologist who applies memory and perception research to the legal system. His research interests include identity verification and face perception, eyewitness memory, police interviewing and forensic science. Richard collaborates with state and federal government, police and emergency services, banks and other financial service providers. He has provided expert evidence in a number of significant court cases, and provides training to judges, lawyers, police and other legal professionals.

DELEGATES

A/Prof Kaye Ballantyne is the Chief Forensic Scientist at the Victoria Police Forensic Services Department in Melbourne, Australia. She is also a member of the Evidence-Based Forensics Initiative.

Thomas Carter is from the Australian Criminal Intelligence Commission.

A/Prof Kim Curby is a cognitive psychologist studying skilled visual performance in face and non-face domains at Macquarie University in Sydney, Australia. She also serves as Deputy Director for Macquarie University's Centre for Elite Performance, Expertise, & Training.

Dr James Dunn is a cognitive psychologist interested in the perceptual and cognitive processes that underlie face identification and expertise at UNSW in Sydney, Australia.

Prof Gary Edmond is a law professor in the Faculty of Law at UNSW in Sydney, Australia where he directs the Program in Expertise, Evidence and Law. He is also the Chair and Founder of the Evidence-Based Forensics Initiative.

Daniel Ferguson is the Team Leader of the Identity Resolution Unit at the Department of Foreign Affairs and Trade's Australian Passport Office in Canberra, Australia.

Jeannine Geach is the Director of Identity and Biometric Futures at the Department of Home Affairs in Canberra, Australia.

Dr Rebecca Heyer is the Group Leader of the Biometrics, Intelligence, Surveillance and Space Division at the Department of Defence, Science and Technology in Adelaide, Australia.

A/Prof Kristy Martire is a cognitive psychologist interested in the development of expertise, processes of evidence evaluation in criminal trials, and improving the communication between experts and lay decision-makers in forensic settings. She is also the Co-Chair of the Evidence-Based Forensics Initiative.

Dr Dana Michalski conducts applied research involving facial comparisons by humans and automated systems at the Department of Defence, Science and Technology in Adelaide, Australia.

Reuben Moreton is a researcher at the Open University studying expertise in applied face matching. He was previously the Senior Facial Image Examiner at the

Metropolitan Police Service and an expert witness in facial identification.

Dr Eilidh Noyes is a cognitive psychologist at the University of Huddersfield, UK who conducts research on human and machine face recognition. She is interested in face recognition for challenging image scenarios and how to achieve the best of human and machine face recognition performance.

Prof Alice O’Toole is a cognitive psychologist at the University of Texas at Dallas, USA interested in human perception, memory, and cognition, and computational approaches to modeling human information processing.

Dr Jonathon Phillips is an Electronic Engineer at the National Institute of Standards and Technology's Information Technology Laboratory in the USA. He is a leading researcher in computer vision, face recognition, biometrics, and forensics and has pioneered competitions to improve technology in these areas.

Dr Kay Ritchie is a cognitive psychologist at the University of Lincoln, UK whose research focuses on human face recognition, improving performance, and public attitudes toward the use of facial recognition technology.

A/Prof Mehera San Roque works in the Faculty of Law at UNSW in Sydney and is involved in research on identification evidence and surveillance technologies aimed at improving the reliability and evaluation of evidence in criminal trials. She is also a member of the Evidence-Based Forensics Initiative.

Dr Clare Sutherland is a cognitive psychologist at the University of Aberdeen, UK interested in facial first impressions and how judgements of faces are influenced by prior knowledge and associated stereotypes.

Cameron Tullberg is a Senior Sergeant at the Victoria Police Forensic Services Department in Melbourne, Australia and is the Contract Manager for the Facial Recognition Team.

A/Prof Jeremy Wilmer is a cognitive psychologist at Wellesley College, USA interested in clinical and non-clinical human variation in cognitive and perceptual abilities to gain insights into their origins, organisation and utility.

GLOSSARY

1-TO-N SEARCH A face matching task performed by facial recognition technology where a ‘probe’ face is used to search a database of known identities for potential matches (see Figure 3). The algorithm will typically return a candidate list of the most similar faces in the database for a human operator to examine. The face matching task performed by the human operator is known as “1-to-N matching”.

1-TO-1 VERIFICATION A face matching task where two faces are compared to decide if they show the same person or two different people. This task can be performed by humans or algorithms.

CANDIDATE LIST A gallery of faces returned by a 1-to-N search. Faces are included in the candidate list if they exceed a threshold of similarity to the probe image (see Figure 3B).

FACE IDENTIFICATION An umbrella term used to describe any task that involves determining a person’s identity from their face. It can include face matching or face recognition memory.

FACE IDENTIFICATION SYSTEM An umbrella term for an organisation’s complete “end-to-end” process of producing identification decisions. Face identification systems can include any combination of humans, algorithms or specialist teams.

FACE MATCHING A type of face identification task where faces (photo, video, live) are simultaneously compared to decide if they show the same person or different people.

FACE MEMORY A type of face identification task where an observer decides if a face has been encountered before. Examples of this task include recognising familiar faces (e.g. family, friends), and when searching for a face in a crowd after memorising faces on a watchlist.

FACIAL EXAMINER/FACIAL FORENSIC EXAMINER Facial examiners are specialist facial image comparison practitioners who may resolve challenging cases and prepare face identification evidence for court. On average, Facial Examiners outperform untrained novices and other groups of practitioners.

FACIAL RECOGNITION TECHNOLOGY/FACIAL RECOGNITION ALGORITHMS Artificial intelligence (AI) systems programmed and trained to make face identification judgements. This can include both 1-to-1 verification and 1-to-N search (see definitions of terms below).

FAMILIAR FACES Faces of people known to an observer. This includes the faces of family, friends and colleagues, but also the faces of celebrities and people we encounter regularly (e.g. barista at a café). Familiarity is developed over multiple, separate encounters.

HYBRID HUMAN-AI EXPERT SYSTEM

A type of end-to-end face identification system that involves decisions made by both human and algorithms. In these systems, human and algorithm decision-makers interact, either by making decisions based on the output of the other or by aggregating independent decisions.

MATCH SCORE THRESHOLD The threshold set by facial recognition technology administrators that determines the approximate false match rate that would be acceptable under operational settings. In 1-to-1 verification tasks, this threshold will determine the level of similarity necessary for the system to deem two images as a match. In 1-to-N search, this threshold will determine the level of similarity necessary for the system to include faces in the candidate list. In both cases, higher thresholds will reduce the number of correct matches but also reduce false matches, while lower thresholds will increase the number of correct matches but also increase false matches.

SUPER-RECOGNISERS Super-recognisers are people with innate superior face recognition ability. Super-recognisers score in the top 1-2% of the population on standardised face identification tests.

UNFAMILIAR FACES Faces of unknown or recently learned people. Unfamiliar faces characterise almost all of face identification decisions made in forensic contexts, as the observer has no previous history with the person they are required to identify.

APPENDIX

A1: WORKSHOP SCHEDULE

Day 1 | Monday 6 January 2020

9:00 – 10:00am	Introduction
10:00 – 11:00am	Three perspectives from psychologists on what it means to be an expert. *Kristy Martire (UNSW), Kim Curby (Macquarie), Alice Towler (UNSW)*
11:00 – 11:30am	Break
11:30 – 1:00pm	Targeted Discussion 1: Individual differences in face perception and recognition. *Jeremy Wilmer (Wellesley College), Romina Palermo (University of Western Australia), Linda Jeffery (University of Western Australia)*
1:00 – 2:00pm	Lunch
2:00 – 3:30pm	Targeted Discussion 2: Face recognition by humans and machines. *Alice O'Toole (University of Texas at Dallas)*
3:30 – 4:00pm	Break
4:00 – 5:30pm	Targeted Discussion 3: Testing face identification experts. *Jonathon Phillips (National Institute of Standards and Technology)*

Day 2 | Tuesday 7 January 2020

9:00 – 9:30am	Recap on Day 1
9:30 – 11:00am	Targeted Discussion 4: Face identification in investigation and evidence. *Gary Edmond (UNSW), Mehera San Roque (UNSW), Kaye Ballantyne (Victoria Police Forensic Services)*
11:00 – 11:30am	Break
11:30 – 12:30pm	Breakout Session 1. *Small group discussions to discuss and reach consensus on main points raised in targeted discussion sessions.*
12:30 – 1:30pm	Lunch
1:30 – 3:30pm	Breakout Session 2. *Small group discussions to discuss and reach consensus on main points raised in targeted discussion sessions.*
3:30 – 4:00pm	Wrap-up and plans for producing digested analysis
5:00 – 7:00pm	Poster session

A2: FIGURE 1 SOLUTION

A	B	A	A	A	B	A	B	A	B
A	A	A	A	A	B	B	B	A	B
B	B	B	A	A	A	B	B	A	A
B	A	B	A	A	B	B	B	B	B