Dr. David White[a], Dr. Alice Towler[a], Prof. Richard I. Kemp[a], Prof. Gary Edmond[a], A/Prof. Mehera San Roque[a], Dr Clare Sutherland[b], Prof. Dame Vicki Bruce, DBE[c] Prof. A. Mike Burton[d] , [A] *UNSW Sydney, Australia,* [B] *University Of Aberdeen, UK,* [C] *Newcastle University, UK,* [D] *University Of York, UK* — Written evidence (NTL0012)

Human oversight of Facial Recognition Technology in forensic applications

Background

1. Our submission relates exclusively to facial recognition technology (FRT). We are an interdisciplinary and international group of cognitive scientists, forensic psychologists, and legal scholars. We write this submission in our capacity as academic experts with decades of experience in the study of facial recognition, as it applies both to human viewers and technology. We have worked with government and police on applied issues relating to face identification in the UK (DW, AT, MB, VB) and Australia (DW, AT, RK, GE), and more recently to improve the way that FRT is implemented and used by staff to protect against identity fraud and investigate crime (DW, AT, RK). Others are experts in the use of image evidence in court in the UK and Australia (GE, MSR). Much of our work on this topic has been published in peer-reviewed academic journals and is freely available to the public.

2. Our submission is based on a recent workshop hosted at UNSW Sydney that focused on human use of FRT. The workshop attracted an international delegation of academic experts, policy makers, law enforcement and FRT practitioners. The resulting white paper is publicly available[1].

Summary of evidence

3. There is a substantial scientific literature on face recognition as it applies both to humans and technology, built on over 50 years of research. This research provides some key facts that are critical to this enquiry.

4. First, FRT requires human oversight via intervention, interpretation and monitoring. This is because in legal and forensic applications, FRT does not 'recognise' faces. Rather, final face identity decisions are made by human operators, who must select faces from 'candidate lists' of potential matches provided by FRT database searches (see para. 18)[2]. Such human oversight of FRT is critical because in forensic applications, FRT is not sufficiently accurate to have high confidence in the veracity of potential matches it identifies (para 7). This is acknowledged in recent updates to the UK Surveillance Camera Code of Practice[3], which state that use of FRT "*should always involve human intervention*

[1] https://socialsciences.org.au/workshop/evaluating-face-identification-expertise-turning-theory-into-practice/

[2] Footnote 1, Pages 7-9; Davies, B. et al (2018). An Evaluation of South Wales Police's Use of Automated Facial Recognition.

before decisions are taken that affect an individual adversely". This 'golden rule' of human oversight has been publicly adopted by the Australian Federal Government[4]: "*Decisions that serve to identify a person will never be made by technology alone*".

5. Second, research has shown that human operators make 50% errors on average when deciding which faces in candidate lists match the search image. This is consistent with research on eye-witness identification – which is known to be unreliable, with well-meaning witnesses often mistakenly identifying innocent suspects[5]. It is also consistent with the fact that this problem persists in tasks with no memory demands – people are commonly wrong when identifying face images they do not have to remember[6]. This problem persists even when the task is performed professionally, for example when checking passports[7].

6. Third, intuition is a poor predictor of accuracy. Despite recognising faces of friends and family in everyday life, we are bad judges of which unfamiliar faces are hard or easy to identify in identification tasks[8]. Most people are poor at this task and yet believe the task is more straightforward than it is[9].

7. Fourth, benchmark tests of algorithm accuracy are conducted in ideal conditions, isolated from normal operational workflow that includes human intervention (paras. 20, 21). Image quality in these tests is unrealistically good due to controlled image capture conditions, which are not possible in criminal investigations and surveillance. Even in ideal testing conditions, errors are common and disproportionally more likely for some demographic groups over others[10]. This is also true of face recognition by human participants[11].

8. Fifth, human decision making is prone to biases[12] whereby errors are influenced by contextual information. These biases can be induced by 'match score' information displayed to operators by FRT[13], which can potentially lead to confident misidentification errors.

[3] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1010815/Surveillance_Camera_Code_of_Practice__update_.pdf

[4] https://www.theguardian.com/technology/2019/sep/29/plan-for-massive-facial-recognition-database-sparks-privacy-concerns

[5] Steblay, N et al. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and human behavior*, *25*(5), 459-473.

[6] E.g. Bruce, V et al. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339-360.

[7] White, D. et al. (2014). Passport officers' errors in face matching. *PLoS ONE, 9*(8), 1-6.

[8] Zhou, X., & Jenkins, R. (2020). Dunning–Kruger effects in face perception. *Cognition, 203*, 104345.

[9] Ritchie, K. L(2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition, 141*, 161-169.

[10] Grother, P. et al. (2019). *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*.

[11] There is a tendency for people to make higher proportions of face recognition errors when recognising faces from other ethnicities (see Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*, 3-35)

[12] Dror, Itiel E., et al. "Letter to the editor—The bias snowball and the bias cascade effects: Two distinct biases that may impact forensic decision making." *Journal of forensic sciences* 62.3 (2017): 832-833.)

9. Sixth, people vary in their ability to recognise faces, and this is a stable cognitive trait that is largely determined by genetic factors[14]. By selecting people who are skilled in face identification to become operators of FRT, and providing them with appropriate training, the probability of error is substantially reduced[15]. Face recognition ability is independent of general IQ and other visual processing[16], but can be reliably measured using tests that target this specific perceptual ability[17].

10. Finally, recent research shows that aggregating face identity judgments made by multiple humans and algorithms produces optimal facial recognition accuracy. Together with recent advances in understanding the variability in face recognition ability from one person to the next, this 'wisdom of crowds' approach provides a simple yet highly effective roadmap for minimising errors associated with the use of FRT[18].

11. In the remainder of this submission we: (i) outline four critical *Knowledge Shortfalls* relating to FRT in forensic applications that should be addressed to provide a basis for policy decisions (para. 12-14); (ii) propose three *Principles for FRT in forensic applications* that are intended to reduce error in human use of FRT and promote effective human oversight (para. 15-17); (iii) provide additional supporting information to assist the enquiry (para. 18-27).

Knowledge shortfalls

12. ***Shortfall 1*. We do not know the accuracy of FRT in realistic applied settings**. There is very little evidence of FRT accuracy in tests that capture the range of real-world operational deployments. Proper end-to-end evaluation must include measurement of accuracy of the total system which includes both FRT *and the people that use FRT*. Such assessments of *system* accuracy are rare. Where operational tests have taken place, in studies of 'live' FRT (defined in para. 18.II)[19], accuracy of human operators has not been reported. This has led to disagreement about how accuracy should be measured and uncertainty about

[13] Howard, J. J. et al. (2020). Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. *PLoS ONE, 15*(8).

[14] Wilmer, J. B. et al. (2010). Human face recognition ability is specific and highly heritable. Proc Nat Academy Sci, 107, 5238-5241. Via; See also Footnote 9

[15] White, D., et al. (2015). Error rates in users of automatic face recognition software. *PLoS ONE, 10*(10), 1-14.

[16] Wilmer, J. B. et al. (2012). Capturing specific abilities as a window into human individuality. *Cognitive Neuropsychology, 29*(5-6), 360-392. Sutherland, C. A. M. et al. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proc Nat Academy Sc, 117*(19), 10218-10224.

[17] White, D. et al. (2021). GFMT2: A psychometric measure of face matching ability. *Behavior Research Methods*.

[18] Phillips et al. (2018). Face recognition accuracy in forensic examiners, super-recognisers and algorithms. *Proc Nat Academy Sci, 115*(24).; Edmond et al.. Facial Recognition and Image Comparison Evidence: Identification by Investigators, Familiars, Experts, Super-recognisers and Algorithms. *Melbourne Uni Law Rev, 45.*

[19] 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. See also Footnote 2

the value of interpretations (para. 19). Given the complexity of these systems in operation, and the wide range of accuracy that is expected in different types of application (see para. 18), devising a standard protocol for measuring accuracy is a significant challenge, and one that requires experts from multiple disciplines (e.g. computer science, cognitive science and psychology, forensic science, policing, law, policy – paras. 25, 26)[20].

13. ***Shortfall 2*****. We do not know how police currently use FRT.** There is a lack of information and understanding relating to how FRT is used in police investigations, despite nearly a decade of police use of FRT (mostly 'retrospective search', see para. 18.I). Lack of clarity limits potential for improvement, auditing, and prohibits public scrutiny of the costs and benefits of FRT. In the US, it recently became apparent that police forces do not keep records or audit the use of FRT, despite active use in many jurisdictions[21]. Freedom of information studies in the US have highlighted examples of improper use that go beyond the intended use of FRT, with potentially profound outcomes in criminal investigations[22]. This points to the importance of maintaining detailed records of how FRT is being used, by whom, and what the outcomes of this use have been.

14. ***Shortfall 3.*** **Potential for FRT to improve forensic science has not been fully explored.** Experts often submit forensic reports in court to support facial image comparisons – for example between CCTV images and reference images from police and other databases, including images sourced from public sites. But the basis of this human expertise has not been clearly articulated in law and has led to errors[23]. Recent scientific studies show that combining human perceptual judgments and FRT match scores in face identity judgments can produce very accurate results[24]. This combination provides a potential route to improving the forensic science of facial image comparison, bringing the forensic science of CCTV identification in line with other forensic science disciplines where algorithms are used to support human judgments[25]. However, use of FRT for this purpose needs to be supported by improved understanding of the strengths and limitations of FRT and forensic science practitioners when performing forensic facial image comparison in realistic conditions, for example where image quality is poor.

Principles for the use of FRT in forensic applications

15. ***Principle 1.*** **Appropriate human oversight and attention to human operators in the design and implementation of facial recognition**

[20] Footnote 1, page 32.

[21] Facial Recognition Technology: Federal Law Enforcement Agencies Should Better Assess Privacy and Other Risks. US Government Accountability Office. Via: https://www.gao.gov/products/gao-21-518; see also https://www.perpetuallineup.org/, https://humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-final-report-2021 (Chapter 9)

[22] https://www.flawedfacedata.com/

[23] Tully, G., & Stockdale, M. (2019). Commentary on: Hak. Evaluation of the Forensic Science Regulator's recommendations regarding image comparison evidence. Forensic science international: Synergy 2019; 1(1). Forensic science international. Synergy, 1, 298-

[24] See Footnote 18, Phillips et al. (2018).

[25] See Footnote 18, Edmond et al. (in press).

systems. Humans are a critical component to ensuring the accuracy of facial recognition decisions made by FRT. Deliberate efforts must be made to ensure that the humans involved in facial recognition decisions are highly skilled, either by targeted recruitment or evidence-based training (see para. 24). Scientifically validated tests of face identification ability can be used to assess suitability of FRT operators and monitor performance over time (see para. 9, footnote 17). In addition, the design of facial recognition systems needs to incorporate appropriate checks and balances to minimise the risk of consequential errors.

16. ***Principle 2.*** **Understanding and disclosure of accuracy, strengths and limitations of face recognition systems, and auditing of their operational use.** Use of FRT in the legal system should be accompanied by transparent disclosure of information relating to the accuracy, strengths, limitations, and operation of this technology. It is also necessary to understand how the technology is being used and to design regulation that ensures appropriate use (para. 23).

17. ***Principle 3.*** **Development of an expert workforce in facial recognition.** If FRT is to be adopted in forensic practice, then new types of expert practitioners and researchers ('meta-experts') are required to design, evaluate, oversee, and explain the resultant face identification systems. Because these systems incorporate human and AI decision making, people with broad expertise in related disciplines are required (paras. 25, 26).

Further supporting information

18. In our submission we distinguish between three different forensic applications of FRT[26] that require different system design considerations. Our submission primarily relates to the most common current use of FRT in forensic applications – one-to-many retrospective database searches (I) – and related applications in live watchlist searches (II). But we also highlight the potential future role for FRT in formalising forensic science practice (III).

I. ***Retrospective database search (one-to-many)***. The most common use of FRT is in police investigations, to enable searching of image databases using facial image evidence. This retrospective database search is often called 'one-to-many' facial recognition. Images that have been gathered during an investigation (e.g. from an ID photo, a CCTV image, or a smartphone) are used to search databases of known people (e.g. mugshots). The human operator using the system will be presented with a 'candidate list' of potential matches to review. This use of FRT also secures systems against identity fraud – for example in passport issuance – by ensuring that applicants do not exist on these systems under several different names.

II. ***Live watchlist search (one-to-many).*** We are aware of several 'Live' facial recognition trials where data streams from CCTV are monitored by facial recognition algorithms in real time, typically checking each face that is detected in the CCTV stream against a 'watchlist' containing persons of

[26] See Footnote 1, pages 7-9 for elaboration

interest[27]. As with retrospective database searching, this function presents human operators with 'candidate lists' of potential matches to review.

III. ***Forensic facial image comparison (one-to-one).*** Law enforcement agencies are often required to compare two images and decide if they show the same person, both in criminal investigations and when submitting forensic evidence to court. Note that while FRT is used to gather intelligence leading to criminal prosecutions and trials, it has not to our knowledge been used to conduct one-to-one comparisons that are submitted as evidence in a court of law. This is incongruous with the higher level of scientific scrutiny applied to technological and human expert performance in one-to-one comparison. Courts currently rely on human experts to make one-to-one comparisons of facial images but the basis of their expertise is often unclear and human judgments are error prone[28]. Evidence shows that errors can be reduced by combining human judgments with match (similarity) scores that are provided by FRT in certain circumstances[29]. There is a need for further research to formalise forensic use of FRT.

19. Accuracy rates for one-to-many searches vary widely, due to differences in test image data, test protocol and how accuracy is measured. This is highlighted in two recent studies commissioned by the London Metropolitan Police Service to measure the operational accuracy of a Live Facial Recognition system[30]. While the MET Police chose to report error-rate as a function of the total number of people that were scanned by the software (i.e. the denominator was every person that passed the camera), giving a 0.1% 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, giving a vastly higher false-positive rate of 80%. Accepted methods for measuring end-to-end accuracy have not been developed and any development would require broad consultation with experts in computer science, behavioural science, law, policing and forensic science.

20. Standard benchmarking tests like those administered by NIST – and quoted by algorithm vendors when asked about the accuracy of their FRT – underestimate the error that would be found in the vast majority of forensic applications. This underestimation is because image quality is often much higher, and databases that are being searched are often much smaller, than those in criminal investigations. Test images are also of 'compliant' subjects that are looking straight at the camera in good lighting and so do not suffer from suboptimal camera angles that are typically encountered in investigations. Algorithm vendors can also be much more certain of the types of images their algorithms will have to process in NIST tests relative to in operational deployment – and prepare their submissions accordingly. Despite these relatively favourable conditions, most algorithms tested in industry benchmarking exercises produce errors at a relatively high rate[14], particularly for some demographic groups compared to others[31].

[27] Footnote 1, Footnote 15

[28] See Footnote 23 & Footnote 18, Edmond et al. (in press).

[29] See Footnote 18, Edmond et al. (in press).

[30] Footnote 19

[31] e.g. Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face Recognition Vendor Test (FRVT). Part 3: Demographic effects*. NIST, U.S. Department of Commerce.

21. In benchmarking tests described above, accuracy is often measured as the probability that a matching face from a given database is returned as the highest ranked possible match. However, because this probability is reduced for low quality imagery used in criminal investigations, it means that human operators must review large lists of images – sometimes over 100 – to ensure they do not miss a match. We tested the face matching accuracy of staff using FRT in their daily work and found errors in 50% of their candidate list review decisions[32], despite displaying lists of just 8 images. Errors indicate that the probability of human operators selecting an innocent person from an FRT candidate list face – a 'false positive' identification (see Footnote 1, page 27) – is alarmingly high.

22. It may be tempting to conclude that because FRT is used to generate investigative leads, and not make definitive judgements of identity, the problem of 'false positives' is not serious. However, broader understanding of the influences of cognitive bias in forensic decision making – and the compounding effects that an error at one stage of an investigation can have on interpretation of subsequent sources of information[33] – suggest that even erroneous investigative leads can lead to serious outcomes such as wrongful arrest[34]. This is especially problematic if other circumstantial evidence means that the gallery is composed of images of plausible suspects – for example those with previous convictions – or even people living in the locality of the offence.

23. The clear implication of the research is that humans provide a critical safeguard against errors resulting from the use of FRT. But humans can also introduce errors if they are not appropriately trained and their abilities have not been assessed. So, human intervention is necessary, but not *sufficient* to ensure proper use. It is therefore paramount that the people tasked with using this technology are aware of the performance profiles of the FRT they are using, i.e. the likelihood the technology will produce errors given varying levels of image quality, for different demographic groups, candidate list sizes, and in different use cases. Globally, there is insufficient public information about who is using FRT, or their training and expertise. It is likely that human operators often do not understand FRT technology and its limitations, given the apparent widespread misuse reported in the US[35].

24. Lab-based tests suggest that some training methods provide benefits to face identification, but these are small in comparison to selection and recruitment[36]. Commercial training courses are often ineffective despite positive reviews from trainees[37], and so it is important that FRT training courses are formally evaluated for their effectiveness.

[32] White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE, 10*(10), 1-14.

[33] Dror, Itiel E., et al. "Letter to the editor—The bias snowball and the bias cascade effects: Two distinct biases that may impact forensic decision making." *Journal of forensic sciences* 62.3 (2017): 832-833.

[34] https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

[35] e.g. Footnote 6

[36] Towler, A. et al (2021). Can face identification ability be trained? Evidence for two routes to expertise. In M. Bindemann (Ed.), *Forensic face matching: Research and practice*: Oxford University Press. Via:

25. Proper implementation of facial recognition systems is therefore more complex than simply purchasing the latest algorithm. While algorithm vendors should do more than simply install the software, they cannot be expected to provide the necessary oversight. This requires 'in-house' testing using images, procedures, and tools representative of the organisation's casework, rather than relying on vendor performance rates or standard benchmark tests. Critically, this testing should consider *both* FRT and human processing accuracy. It requires specialist staff to be accountable for the whole system, including algorithms, workflow design, delegation of tasks to humans etc. Staff must have expertise to measure accuracy and report risks associated with deployment of algorithms for specific uses.

26. Similarly, the cost of implementing a facial recognition system far exceeds the cost of purchasing and installing FRT in IT systems and incorporating it into workflow. Budget allocations for FRT must include funds for ensuring appropriate use and oversight of the system, including regular testing of facial recognition system accuracy in operational deployment, selection of human operators, staff training, career development of FRT specialist teams, routine monitoring and auditing of FRT use, user experience design etc.

27. To conclude, properly implemented facial recognition systems can provide benefits in forensic settings – both in criminal investigations and in ensuring courts have access to the best evidence available. But it is important that these systems are designed with current understanding of human and machine facial recognition abilities in mind. Critically, accuracy of these systems must be tested and monitored in conditions that effectively capture their operational deployment, using test protocols that measure errors made both by FRT, and by the people that use this technology.

3 September 2021

[37] Towler, A. et al. (2019). Do professional facial image comparison training courses work? *PLoS ONE, 14*(2), e0211037.