FACIAL RECOGNITION AND IMAGE COMPARISON EVIDENCE: IDENTIFICATION BY INVESTIGATORS, FAMILIARS, EXPERTS, SUPER-RECOGNISERS AND ALGORITHMS

GARY EDMOND,[*] DAVID WHITE,[**] ALICE TOWLER,[†] MEHERA SAN ROQUE[‡] AND RICHARD KEMP[‡‡]

Drawing upon decades of scientific research on face perception, recognition and comparison, this article explains why conventional legal approaches to the interpretation of images (eg from CCTV) to assist with identification are misguided. The article reviews Australian rules and jurisprudence on expert and lay opinion evidence. It also summarises relevant scientific research, including emerging research on face matching by humans (including super-recognisers) and algorithms. We then explain how legal traditions, and the interpretation of rules and procedures, have developed with limited attention to what is known about the abilities and vulnerabilities of humans, algorithms and new types of hybrid systems. Drawing upon scientific research, the article explains the need for courts to develop rules and procedures that attend to evidence of validity, reliability and performance — ie proof of actual proficiency and levels of accuracy. It also explains why we should resist the temptation to admit investigators' opinions about the identity of offenders, and why leaving images to the jury introduces unrecognised risks by virtue of the surprisingly error-prone performance of ordinary persons and the highly suggestive (or biasing) way in which comparisons are made in criminal proceedings. The article recommends using images in ways that incorporate scientific knowledge and advance fundamental criminal justice values.

CONTENTS

[*] Professor, Faculty of Law and Justice, University of New South Wales; Professor (fractional), School of Law, Northumbria University; Chair, Evidence-Based Forensics Initiative. This research was supported by the ARC (LP16010000). Thanks to Jason Chin and reviewers for comments.

[**] ARC Future Fellow, Senior Lecturer, School of Psychology, University of New South Wales.

[†] ARC DECRA Fellow, School of Psychology, University of New South Wales.

[‡] Associate Professor, Faculty of Law and Justice, University of New South Wales.

[‡‡] Professor, School of Psychology, University of New South Wales.

I Improving Legal Practice with Scientific Research

This article is about the use of images in criminal proceedings.[1] It considers how courts should engage with the interpretation of images to assist with identification and, inexorably, the role of scientific research in the admission, presentation and evaluation of this evidence. The article provides an overview of scientific research on the use of images for the purpose of identifying persons of interest ('POIs'). This follows a review of the piecemeal Australian jurisprudence pertaining to the admission and interpretation of image evidence.[2]

[1] It builds upon a series of law-related articles we began publishing more than a decade ago: see, eg, Gary Edmond et al, 'Law's Looking Glass: Expert Identification Evidence Derived from Photographic and Video Images' (2009) 20(3) *Current Issues in Criminal Justice* 337 ('Law's Looking Glass'); Gary Edmond et al, '*Atkins v The Emperor*: The "Cautious" Use of Unreliable "Expert" Opinion' (2010) 14(2) *International Journal of Evidence and Proof* 146 ('*Atkins v The Emperor*'); Gary Edmond, Josh P Davis and Tim Valentine, 'Expert Analysis: Facial Image Comparison' in Tim Valentine and Josh P Davis (eds), *Forensic Facial Identification from Eyewitnesses, Composites and CCTV* (Wiley Blackwell, 2015) 239. It also builds on work from Paul Bogan and Andrew Roberts, *Identification: Investigation, Trial and Scientific Evidence* (Jordan Publishing, 2nd ed, 2011); Josh P Davis and Tim Valentine (eds), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (Wiley Blackwell, 2015).

[2] See below Part II.

Juxtaposed, these two reviews palpably demonstrate the limits of legal concepts and prevailing practice. It is our contention that recourse to scientific knowledge can help to ensure the admission of opinions that are reliable (and actually expert) and likely to enhance factual accuracy, efficiency and the fairness of criminal justice processes.

Confronted with the rapid expansion in the availability of images, from the turn of the millennium Australian courts imposed epistemologically arbitrary constraints on the use of images to assist with the identification of POIs.[3] These restrictions appear to have been based on judicial anxiety about the value of the opinions but are blind to the actual abilities or accuracy of purported identification experts. Drawing upon mainstream scientific knowledge, we explain what courts should be looking for in order to identify: those who possess actual expertise; the scope and limits of their abilities; and how their opinions ought to be expressed. We also draw upon this knowledge base to consider how we might respond to emerging forms of expertise and expert systems that do not necessarily sit comfortably within conventional legal categories and conceptualisations — such as the opinions of *super-recognisers*,[4] the outputs of face matching algorithms and even combinations of humans and/or algorithms (ie hybrid systems).[5] This article offers an evidence-based approach to the interpretation of images — that is, it is concerned with scientific *knowledge*. Therefore, the article is conspicuously sensitive to accuracy and presenting evidence (whether an opinion or the output of an algorithm) in a manner that embodies its value so that it is susceptible of rational evaluation.[6]

In this article we refer to those whose abilities at face comparison have *not* been formally established — ie demonstrated in controlled conditions where the correct answer is known — as *purported experts*. We call them 'purported'

[3] See below Part II(A). See also Richard I Kemp, Gary Edmond and David White, 'A Proposed Solution to the Problem of Identifying People from CCTV and Other Images' in Andrew M Smith, Michael P Toglia and James Michael Lampinen (eds), *Methods, Measures, and Theories in Eyewitness Identification Tasks* (Routledge, 2021) 13, 18.

[4] Super-recognisers are individuals at the extreme end of the distribution of natural ability in face matching (and memory). They are discussed below in Part III(B)(3).

[5] See Andrea Roth, 'Machine Testimony' (2017) 126(7) *Yale Law Journal* 1972, 1972, 1998–9; Kemp, Edmond and White (n 3) 25–6. See also *R (Bridges) v Chief Constable of South Wales Police* [2020] 1 WLR 5037, 5045 [15]–[16] (Sir Terence Etherton MR, Dame Victoria Sharp P and Singh LJ).

[6] Gary Edmond, 'Forensic Science Evidence and the Conditions for Rational (Jury) Evaluation' (2015) 39(1) *Melbourne University Law Review* 77, 125–7. This commitment to 'rationality' is drawn from ss 55–6 of the *Uniform Evidence Law* and the definition of 'probative value' in s 3. The *Uniform Evidence Law* is the collection of the *Evidence Act 1995* (Cth); *Evidence Act 2011* (ACT); *Evidence Act 2004* (Norfolk Island); *Evidence Act 1995* (NSW); *Evidence (National Uniform Legislation) Act 2011* (NT); *Evidence Act 2001* (Tas); *Evidence Act 2008* (Vic).

because we do not know if they are actually expert at the specific task, namely identifying a POI or describing their features.[7] For more than a decade (up to 2014), Australian courts allowed purported experts to identify POIs or features said to be shared between a POI and a defendant.[8] Purported experts were not required to provide information about their abilities, the degree to which facial features are correlated, or how patterns of correlations vary (ie diagnosticity).[9] Evaluation of the opinions of purported experts was left to the exigencies of the trial. Reception was moderated by trial safeguards (eg cross-examination and judicial instructions) that were not always informed by contemporary scientific knowledge.[10] Eventually, following *Honeysett v The Queen* ('*Honeysett*'), (some) purported experts were prevented from proffering opinions, though the extent of that proscription remains uncertain.[11] Overall, Australian judges erected an admissibility regime that directs limited attention to the epistemic value of images or the serious risks posed by their interpretation.[12]

This article is intended to encourage courts to reconsider their position(s). Courts, we contend, should be more sceptical consumers of facial recognition and image comparison evidence. Guided by scientific research, courts can and

[7] Images of known persons, when used for comparison purposes, are often described as reference images: see Gary Edmond, 'A Closer Look at *Honeysett*: Enhancing Our Forensic Science and Medicine Jurisprudence' (2015) 17(2) *Flinders Law Journal* 287, 293–4 ('A Closer Look at *Honeysett*'). These are analogous to the fingerprint records on file, which are used to *identify* latent prints from a crime scene.

[8] See *Honeysett v The Queen* (2014) 253 CLR 122, 138 [45]–[46] (French CJ, Kiefel, Bell, Gageler and Keane JJ) ('*Honeysett*'). See also Kemp, Edmond and White (n 3) 18.

[9] See, eg, *R v Tang* (2006) 65 NSWLR 681, 709 [120], 712 [135] (Spigelman CJ) ('*Tang*'). Contrast other scientific comparison procedures, such as DNA profiling, where statistical information is generated in the context of known frequencies and patterns of association: Gary Edmond, 'What Lawyers Should Know about the Forensic "Sciences"' (2015) 36(1) *Adelaide Law Review* 33, 38 ('What Lawyers Should Know').

[10] See *Tang* (n 9) 691 [42]–[43] (Spigelman CJ).

[11] *Honeysett* (n 8) 138 [45]–[46] (French CJ, Kiefel, Bell, Gageler and Keane JJ). See Edmond, 'A Closer Look at *Honeysett*' (n 7) 299–300. The present article is primarily, though not exclusively, concerned with *Uniform Evidence Law* (n 6) jurisdictions. Though, the scientific evidence and criticisms apply Australia-wide.

[12] Australia is not alone. Historically, the jurisdiction of England and Wales was even more liberal in its admissibility practice. There were few constraints on expert opinion in criminal proceedings and those admitted as experts were free to categorically identify POIs in images: see, eg, *A-G's Reference (No 2 of 2002)* [2003] 1 Cr App R 21, 323 [2], 327–8 [19]–[21] (Rose LJ for the Court); Gary Edmond and Natalie Wortley, 'Interpreting Image Evidence: Facial Mapping, Police Familiars and Super-Recognisers in England and Australia' (2016) 3(2) *Journal of International and Comparative Law* 473, 479. Recent reforms to criminal procedure rules in England may have tightened admissibility in England: see *Criminal Procedure Rules 2015* (UK) pt 19; Michael Stockdale and Adam Jackson, 'Expert Evidence in Criminal Proceedings: Current Challenges and Opportunities' (2016) 80(5) *Journal of Criminal Law* 344, 352–3.

should be more accommodating of *some* types of opinion evidence, but less accommodating of others. They should possess effective means of distinguishing between different types of direct and indirect witnesses, determining which indirect witnesses have genuine abilities, and gauging the probative value of both opinions and the outputs of face matching algorithms.[13] They should also be more attentive to scientific research on unconscious bias and its deleterious impacts on perception and interpretation.[14] Our analysis begins with a synoptic overview of the main developments in the Australian jurisprudence over the last two decades, including the expanding reliance on common law categories such as the 'ad hoc' expert.[15] In Part III we have assembled and synthesised scientific research on unfamiliar face comparison to provide the foundations for a more principled approach to the use of images. Drawing on these reviews, Part IV raises a series of considerations (and makes a few recommendations) that ought to inform legal reliance on images used as evidence of identity in criminal proceedings.

II Legal Interpretations of Images and Experts

Now ubiquitous, images can fulfil a range of evidentiary functions in investigations and prosecutions. They can be used to identify persons, to track movements and to determine what was done, when and by whom.[16] All of these uses involve interpretations that may vary from the easy and mundane to those which are extremely difficult, contested, and error-prone. The value of images as evidence depends on their use (eg the type of interpretation), the quantity and quality of the images, as well as the abilities of those (whether individuals or algorithms or hybrid systems) interpreting them — see Part III.

In this article we are concerned with the use of images to assist with the identification of POIs — frequently those involved in criminal activity. This can be based on *recognition* (from memory and some degree of familiarity) or the *comparison* of a POI in images with reference images (usually of a person whose

[13] Algorithms are already in use, especially by investigators, and their expanded use in investigations would seem inevitable: see below Part III(B)(4).

[14] See generally Gary Edmond et al, 'Thinking Forensics: Cognitive Science for Forensic Practitioners' (2017) 57(2) *Science & Justice* 144.

[15] Readers from non-Australian jurisdictions might jump to Parts III–IV.

[16] They can also be used to assist eyewitnesses, though the focus of this article is identification by those who rely on images but who did not *directly* perceive the person or events, which may include the jury.

identity is known).[17] We are primarily concerned with identification by strangers based on comparisons — sometimes described as unfamiliar face matching or forensic image comparison.[18] That is, comparisons performed by those with little to no familiarity with the suspect/defendant (prior to an investigation).[19]

A Smith, Morgan, Honeysett *and the Limits of s 79 Jurisprudence*

Our review begins with *Smith v The Queen* ('*Smith*')[20] — the High Court's first attempt to respond to some of the implications of the dramatic expansion in the availability of images.[21] During the trial, two police officers with some limited exposure to Mundarrah Smith purported to identify him in low-quality images of an armed robbery — see Figure 1.[22] On appeal, the High Court deemed this evidence inadmissible.[23] According to the majority, the police

[17] These are ideal types and in practice may overlap. This article is primarily concerned with those who are *not* direct witnesses. Sections 113–16 of the *Uniform Evidence Law* (n 6) cover only those who are giving evidence based on direct perception, and does not regulate image comparison or feature descriptions: see Mehera San Roque, 'Updating Beliefs: Rethinking the Regulation of Identification Evidence under the *UEL*' in Andrew Roberts and Jeremy Gans (eds), *Critical Perspectives on the Uniform Evidence Law* (Federation Press, 2017) 195, 195.

[18] Much of the discussion, along with the need for rigorous research, applies to somatic comparison as well as the comparison of other things in images, including clothes and shoes, tattoos, possessions (eg weapons), as well as movement and posture (eg gait): see, eg, *R v Abbey* (2009) 246 CCC (3d) 301, 312–13 [34]–[37] (Doherty JA for the Court) (Ontario Court of Appeal) and the subsequent appeal in *R v Abbey* (2017) 350 CCC (3d) 102, 105–6 [3]–[5] (Laskin JA) (Ontario Court of Appeal); *Meade v The Queen* [2015] VSCA 171, [148]–[154] (Maxwell P, Redlich and Whelan JJA); *Otway v The Queen* [2011] EWCA Crim 3, [9]–[16] (Pitchford LJ for the Court) ('*Otway*'); *R v Aitken* [2012] BCCA 134, [80] (Hall J, Finch CJ agreeing at [104], Hinkson J agreeing at [104]) ('*Aitken*'). See also Gary Edmond and Emma Cunliffe, 'Cinderella Story: The Social Production of a Forensic "Science"' (2016) 106(2) *Journal of Criminal Law and Criminology* 219, 232–3, 245–50.

[19] Though, we will say a few things about identification by familiars, that is those whose familiarity with the suspect is acquired through various real-world interactions, and contrast it with investigators who are said to become 'familiar' through the course of the investigation — so-called ad hoc experts or lay witnesses, whose opinions might be admissible via *Uniform Evidence Law* (n 6) s 78 or s 79(1).

[20] (2001) 206 CLR 650 ('*Smith*').

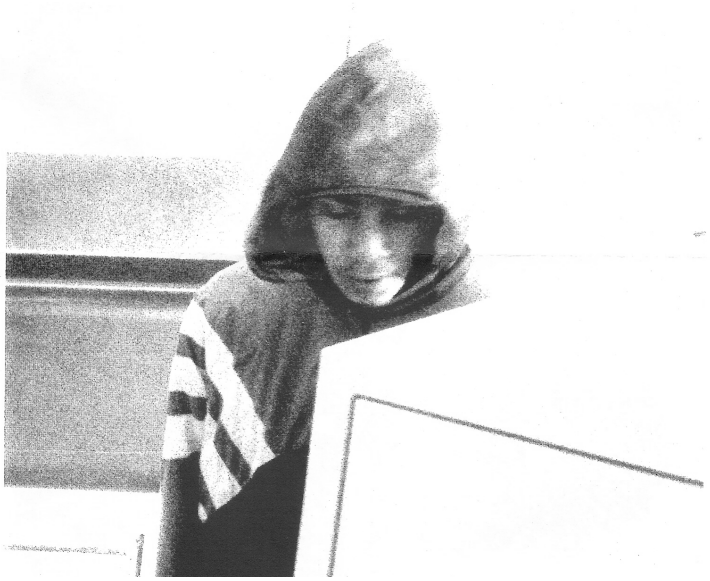
[21] See Edmond et al, 'Law's Looking Glass' (n 1) 377.

[22] *Smith* (n 20) 653 [5] (Gleeson CJ, Gaudron, Gummow and Hayne JJ). In all of the cases in this article that involved CCTV images, the security cameras recorded more than a single frame: *Honeysett* (n 8) 125–6 [1] (French CJ, Kiefel, Bell, Gageler and Keane JJ); *Tang* (n 9) 685 [16] (Spigelman CJ); *Morgan v The Queen* (2011) 215 A Crim R 33, 44 [72] (Hidden J) ('*Morgan*'). The images reproduced here were often relied upon at trial or seem to be understood as among the best of the individual images for presentation at trial.

[23] *Smith* (n 20) 656 [12] (Gleeson CJ, Gaudron, Gummow and Hayne JJ).

officers were not entitled to identify Smith as the POI in the bank robbery because each offered nothing beyond what a jury could bring to the comparison.[24] During the course of the trial, members of the jury would have more exposure to the defendant than either of the police officers obtained during their previous dealings.[25] For the majority, the opinions (ie interpretations) of the police officers would add nothing to the jury comparisons.[26] They could not 'rationally affect (directly or indirectly) the assessment of the probability of the existence of a fact in issue in the proceeding'.[27] They were, by definition, irrelevant.[28]

Figure 1: The bank robber alleged to have been Mundarrah Smith



[24] Ibid 655–6 [11]–[12]. The police officers, who do not appear to have been involved directly in the investigation, had each spent some time with Smith: at 653 [5]. See also Katherine Biber, 'The Hooded Bandit: Aboriginality, Photography and Criminality in Smith v The Queen' (2002) 13(3) *Current Issues in Criminal Justice* 286, 287.

[25] *Smith* (n 20) 654–5 [9] (Gleeson CJ, Gaudron, Gummow and Hayne JJ). Consider Canadian jurisprudence in *R v Leaney* [1989] 2 SCR 393, 409 (Wilson J) and *R v Anderson* [2005] BCSC 1346, [16]–[32] (Smith J).

[26] *Smith* (n 20) 655 [11] (Gleeson CJ, Gaudron, Gummow and Hayne JJ).

[27] Ibid 655–6 [10]–[12]. See also *Uniform Evidence Law* (n 6) s 55(1).

[28] *Uniform Evidence Law* (n 6) s 55(1). The majority found the police officer's evidence irrelevant and so did not address the interpretations: *Smith* (n 20) 656 [12].

An exception — the *Smith* caveat — was available where witnesses held some non-trivial advantage over the trier of fact.[29] This might arise where the appearance of the defendant had changed, or movement (eg gait) was significant and they had been exposed to these in ways that gave them a distinct advantage over the trier of fact.[30] The majority explained:

> In other cases, the evidence of identification will be relevant because it goes to an issue about the presence or absence of some identifying feature other than one apparent from observing the accused on trial and the photograph which is said to depict the accused. Thus, if it is suggested that the appearance of the accused, at trial, differs in some significant way from the accused's appearance at the time of the offence, evidence from someone who knew how the accused looked at the time of the offence, that the picture depicted the accused as he or she appeared at *that* time, would not be irrelevant. Or if it is suggested that there is some distinctive feature revealed by the photographs (as, for example, a manner of walking) which would not be apparent to the jury in court, evidence both of that fact and the witness's conclusion of identity would not be irrelevant.[31]

Justice Kirby adopted a different course. He was unwilling to unilaterally invoke (ir)relevance in the High Court when it was not relied on by the parties or judges during the trial and appeals.[32] Rather than treat the police officers' impressions as irrelevant, for Kirby J there was no exception to the exclusionary opinion rule[33] that would render them admissible.[34] The opinions were not 'based on what [the officers] saw, heard or otherwise perceived about a matter or event' and so were not admissible under the exception for lay opinion in

[29] *Smith* (n 20) 656–7 [15]–[17] (Gleeson CJ, Gaudron, Gummow and Hayne JJ).

[30] Consider *Evans v The Queen* (2007) 235 CLR 521, 569–70 [183]–[184] (Heydon J). See also *R v Crupi (Ruling No 1)* [2020] VSC 654, [1], [13]–[17], [87]–[88] (Beale J).

[31] *Smith* (n 20) 656–7 [15] (Gleeson CJ, Gaudron, Gummow and Hayne JJ) (emphasis in original) (citations omitted). That reasoning was applied by Ipp JA in *Li v The Queen* (2003) 139 A Crim R 281, 294–5 [103]–[113] ('*Li*'), in relation to a detective's testimony on the accused's appearance, clothes, posture and vehicle.

[32] *Smith* (n 20) 657–8 [20], 663–4 [41] (Kirby J). Cf *R v Smith* (1999) 47 NSWLR 419, 423–4 [23]–[24] (Sheller JA), where the police officers' evidence was treated as recognition evidence: that is, as a species of *fact*, such that *Uniform Evidence Law* (n 6) s 76 was said not to apply. This possibility is raised by Basten JA as one of three possible admissibility routes in his consideration of voice identification evidence in *Nguyen v The Queen* (2017) 264 A Crim R 405, 409–10 [13]–[15] ('*Nguyen*'). The other two are the exceptions provided by *Uniform Evidence Law* (n 6) ss 78, 79(1): *Nguyen* (n 32) 410 [14]–[15] (Basten JA).

[33] *Uniform Evidence Law* (n 6) s 76.

[34] *Smith* (n 20) 670–1 [64] (Kirby J). Ordinarily, opinions adduced 'to prove the existence of a fact about the existence of which the opinion was expressed' are not admissible: *Uniform Evidence Law* (n 6) s 76. Interpretations of the identity of POIs in images are opinion caught by this rule: see *Smith* (n 20) 669 [58] (Kirby J).

s 78(a) of the *Uniform Evidence Law*.[35] Nor were the opinions based on 'specialised knowledge' — the police officers had none — and so they were not admissible under the exception for expert opinion in s 79(1).[36] Thus, all of the judges agreed that the police officers' opinions were inadmissible, but Kirby J's reasoning offered a potential solution to the majority's proscription by drawing attention to the possibility of admission if the conditions of s 79 could be satisfied.[37]

Indirectly, then, *Smith* encouraged investigators and prosecutors to call upon a variety of witnesses presented as experts, sometimes characterised as facial mappers or face and body mappers, to provide opinions about the identity of the POI in images based on 'training, study or experience' that was presented as apposite.[38] These individuals often possessed formal qualifications and/or experience in anatomy (and other domains) that were *considered by courts* to be capable of grounding an ability to interpret images to assist with identification.[39] As we shall see, when considering admissibility under s 79(1), lawyers and judges were not particularly attentive to the fact that the individuals proffering opinions were almost always experts in domains (or 'fields') not centrally concerned with image comparison or identification.[40] It does not

[35] *Smith* (n 20) 669–70 [60] (Kirby J). The 'matter or event' is the bank robbery and the police officers did not witness it. The matter or event is not the video of the robbery or the identity of the robbers. Historically, lay opinion has been limited to those who directly perceived a matter or event. See also *Lithgow City Council v Jackson* (2011) 244 CLR 352, 372 [46] (French CJ, Heydon and Bell JJ) ('*Lithgow*'). The views expressed in Law Reform Commission, *Evidence* (Interim Report No 26, 1985) vol 1, 410–11 [739]–[740] ('*Evidence Interim Report*') are consistent with confining the application of *Uniform Evidence Law* (n 6) s 78 to eyewitnesses (only). See also the discussion in Australian Law Reform Commission, *Uniform Evidence Law* (Report No 102, December 2005) 282–5 [9.11]–[9.24] ('*Uniform Evidence Law Report*') on whether *Smith* (n 20) gave rise to a need to amend *Uniform Evidence Law* (n 6) s 78 (or the rules in pt 3.3) to enable evidence to be given by investigators.

[36] *Smith* (n 20) 669 [59] (Kirby J). Section 79(1) of the *Uniform Evidence Law* (n 6) states:

> If a person has specialised knowledge based on the person's training, study or experience, the opinion rule does not apply to evidence of an opinion of that person that is wholly or substantially based on that knowledge.

[37] *Uniform Evidence Law Report* (n 35) 284–5 [9.22]–[9.24] fails to address this aspect of the decision in *Smith* (n 20).

[38] *Uniform Evidence Law* (n 6) s 79. We say 'presented' because their abilities had not been demonstrated through processes of evaluation.

[39] See, eg, *R v Alrekabi* (2007) 4 DCLR(NSW) 292, 295 [26]–[27], 297 [36] (Knox DCJ) ('*Alrekabi*'). Purported experts may have formal qualifications in adjacent domains, such as physical anthropology, military surveillance, IT, medical art, photography and so on.

[40] In *Dasreef Pty Ltd v Hawchar* (2011) 243 CLR 588, 604 [37] (French CJ, Gummow, Hayne, Crennan, Kiefel and Bell JJ) ('*Dasreef*'), the High Court provided a 'shortcut' for some types of expert opinion, particularly those in regular use. The majority explained that some expert opinions 'will require little explicit articulation or amplification once the witness has described

follow, for example, that a person with formal qualifications and experience in anatomy, or even facial reconstruction from skeletal remains, will be better than lay persons (eg jurors) at comparing faces, or discerning facial or body features or movement, in images.[41] Expertise cannot be *assumed* to be transferable and knowledge of body parts (and Latin nomenclature) may have no correspondence with interpretive ability.[42] In the admissibility decisions and appeals following *Smith* there is little apparent interest in the performance (ie ability and accuracy) of the purported experts admitted as expert witnesses.[43] The issue of 'specialised knowledge' was considered in only a few instances.[44] The most influential of these decisions was also the least satisfactory.

In *R v Tang* ('*Tang*'), the value of 'expert' interpretations came into question when a challenge to the admissibility of an anatomist's opinions about the identity of a defendant in an armed robbery was raised on appeal.[45] The anatomist testified that the POI in security images and Tang were 'one and the same' — see Figure 2 below.[46] This was based on her comparison of both faces and bodies:

> Her evidence included a comparison between the two sets of material [ie the images of the robbery and reference images of Tang in Figure 2] in terms of identifying similarities such as the thickness of the lips, the existence of a dimple on the chin, the fact that each chin was 'squarish' and the jaw structure 'angular'. She

his or her qualifications and experience, and has identified the subject matter about which the opinion is proffered'. See also Gary Edmond and Kristy Martire, 'Knowing Experts? Section 79, Forensic Science Evidence and the Limits of "Training, Study or Experience"' in Andrew Roberts and Jeremy Gans (eds), *Critical Perspectives on the Uniform Evidence Law* (Federation Press, 2017) 80, 87–8 ('Knowing Experts'). For a review of the consequences of these sorts of shortcuts, see Gary Edmond, 'Latent Science: A History of Challenges to Fingerprint Evidence in Australia' (2019) 38(2) *University of Queensland Law Journal* 301, 337–8.

41 Kristy A Martire and Gary Edmond, 'Rethinking Expert Opinion Evidence' (2017) 40(3) *Melbourne University Law Review* 967, 984–7.

42 Expertise, in the sense of an enhanced ability over ordinary persons (following *Smith* (n 20) 654–5 [9] (Gleeson CJ, Gaudron, Gummow and Hayne JJ), 669 [59] (Kirby J)), is an empirical question. Is this person better at some specific task than ordinary persons, and how do we know? (In some instances the person might just know more, but this is not the case with claimed abilities — such as the ability to accurately interpret images.)

43 See *R v Jung* [2006] NSWSC 658, [63]–[66], [87] (Hall J); *Alrekabi* (n 39) 299 [42]–[43], 302–3 [68] (Knox DCJ); *R v Dastagir* (2013) 118 SASR 83, 95 [63]–[68] (Kourakis CJ, Vanstone and Stanley JJ) ('*Dastagir*'); *Morgan* (n 22) 59–60 [138]–[139] (Hidden J).

44 It is raised, and even defined, but not meaningfully applied in *Tang* (n 9) 712–13 [137]–[140] (Spigelman CJ); *Morgan* (n 22) 59–61 [135]–[145] (Hidden J); *Honeysett* (n 8) 131–2 [22]–[23], 137 [42] (French CJ, Kiefel, Bell, Gageler and Keane JJ).

45 *Tang* (n 9) 683–4 [7]–[8] (Spigelman CJ).

46 Ibid 688 [28] (Spigelman CJ). The anatomist's conclusion was expressed by reference to a similarity scale: at 688 [30].

> also identified, in accordance with her classification that was before the jury, that the two facial forms were 'pentagonal', that the facial height was 'medium', and there was a 'lateral projection' of the cheekbones, meaning a projection to the side. She referred to a wide chin and visible lips seam, a slight projection of the ears and what is described as a 'mezzo-cranic head shape', meaning a short, broad and high head with a flattened occipital region, being the back of the head. She also identified a number of the features in both sets of material as characteristics of an Asian person. She referred to similarities in terms of the 'upright posture of the upper torso'. Dr Sutisno also identified three 'unique identifiers': the lips, the wide square chin with a dimple and the posture.[47]

The New South Wales Court of Criminal Appeal ('NSWCCA') was persuaded about the existence of a 'field' of facial comparison, but not one concerned with comparison of the body:[48]

> The detailed knowledge of anatomy which Dr Sutisno unquestionably had, together with her training, research and experience in the course of facial reconstruction supports her evidence of facial characteristics.
>
> Nothing was presented to the Court which indicates, in any way, that Dr Sutisno's extension from facial to body mapping, with respect to matters of posture, has anything like that level of background and support.[49]

[47] Ibid 687–8 [25].

[48] See ibid 712 [135]–[136] (Spigelman CJ). Note the persistence of common law concepts, such as 'field', which may shift attention from the ability to do some task (such as accurately compare persons in images) and lead to concern with potentially less valuable information such as study and experience (in, for example, anatomy): see at 713 [142] (Spigelman CJ). Of interest, in *Tang* (n 9) 713 [142] (Spigelman CJ, Simpson J agreeing at 716 [159], Adams J agreeing at 716 [160]) the Court relied upon support for spectrographic voice comparison in *R v Gilmore* [1977] 2 NSWLR 935, 937–8 (Street CJ) ('*Gilmore*'), even though that method was criticised following scientific review by the National Academy of Sciences: see Committee on Evaluation of Sound Spectrograms, National Research Council, *On the Theory and Practice of Voice Identification* (Report, National Academy of Sciences, 1979) 2. According to Spigelman CJ in *Tang* (n 9) 713 [142], citing *Gilmore* (n 48) 937–8 (Street CJ):

> There was nothing in the present case remotely like the evidence before this Court when it accepted voice identification as a field of expert study or knowledge which could support evidence in the form that the two voices were the same.

See also the considerable detail explaining the new technique in *R v McHardie* [1983] 2 NSWLR 733, 755–63 (Begg, Lee and Cantor JJ).

[49] *Tang* (n 9) 712 [135]–[136] (Spigelman CJ).

The reasons for the distinction are not particularly clear. It is not a distinction supported by scientific research.[50]

Figure 2: Copy of an exhibit (photo board) produced by the anatomist in Tang*. The POI in an armed robbery is juxtaposed with a reference image of Tang. Note the deliberate orientation of the face in the inset marked '5'.*

[50] A number of sources are cited in the *Tang* (n 9) decision: see, eg, at 699–700 [69]–[72] (Spigelman CJ), citing Mehmet Yasar Iscan, 'Introduction of Techniques for Photographic Comparison: Potential and Problems' in Mehmet Yasar Iscan and Richard P Helmer (eds), *Forensic Analysis of the Skull: Craniofacial Analysis, Reconstruction, and Identification* (John Wiley & Sons, 1993) 57, 57–60; GJR Maat, 'The Positioning and Magnification of Faces and Skulls for Photographic Superimposition' (1989) 41(3) *Forensic Science International* 225; Michael C Bromby, 'At Face Value?' (2003) 153 *New Law Journal* 302, 302–3. *Tang* (n 9) also contains several opaque references: see, eg, at 699 [70] (Spigelman CJ), citing RA Halberstein, 'The Application of Anthropometric Indices in Forensic Photography: Three Case Studies' (2001) 46(6) *Journal of Forensic Sciences* 1438; Glenn Porter and Greg Doran, 'An Anatomical and Photographic Technique for Forensic Facial Identification' (2000) 114(2) *Forensic Science International* 97, 97. None of these articles reports on formal scientific evaluation of face comparison. Iscan (n 50) pertains to skulls, Bromby (n 50) is a short article in a legal periodical, and the process of superimposition, discussed in Maat (n 50), has been criticised and largely abandoned: see Glenn Porter, 'The Reliability of CCTV Images as Forensic Evidence' (PhD Thesis, University of Western Sydney, 2011) 126. These materials are not relied upon to identify 'knowledge' or to support qualified admission. The reference to Porter and Doran (n 50), which is not properly cited in the decision, is an article proposing a method that might improve the accuracy of interpreting features in images: at 97.

Operating in the *Uniform Evidence Law* tradition, where the opinion rule (s 76) covers the field — excluding all opinions adduced 'to prove the existence of a fact about the existence of which the opinion was expressed' unless there is an express exception — the NSWCCA nonetheless unhelpfully drew upon the common law conception of the 'ad hoc expert'.[51] This was presented as consistent with s 79(1) and able to support the admission of an opinion about apparent similarities:

> The identification of points of similarity by Dr Sutisno was based on her skill and training, particularly with respect to facial anatomy. It was also based on her experience with conducting such comparisons on a number of other occasions. Indeed, it could be supported by the experience gained with respect to the videotape itself through the course of multiple viewing [sic], detailed selection, identification and magnification of images. By this process she had become what is sometimes referred to as an 'ad hoc expert'.[52]

'Ad hoc expert' is a legal category that privileges the interpretations of individuals — usually investigators or those working with them — who are said to have obtained an advantage over the trier of fact through repeated exposure to some stimuli (typically some kind of recording).[53] The concept was originally applied to opinions about the *content* of audio recordings — specifically the words allegedly spoken and transcribed[54] — on the basis of repeated listening.[55] Over

[51] *Tang* (n 9) 709 [120] (Spigelman CJ). Note that this alternative avenue for the admission of opinions had been recognised in earlier cases, such as *Li* (n 31) 289 [57] (Ipp JA). See Gary Edmond and Mehera San Roque, 'Quasi-Justice: Ad Hoc Expertise and Identification Evidence' (2009) 33(1) *Criminal Law Journal* 8, 11–14, 17–18. Somewhat remarkably, in *Nguyen* (n 32), Basten JA contends that *Uniform Evidence Law* (n 6) s 76 does not exclude opinions that were otherwise admissible at common law: *Nguyen* (n 51) 413 [27].

[52] *Tang* (n 9) 709 [120] (Spigelman CJ). The justification in this case seems to rely on anatomical 'skill and training', experience performing similar comparisons in other investigations, and repeated viewing of the images in this case. None of this seems to be 'knowledge'.

[53] Edmond and San Roque (n 51) 8.

[54] See the limited scope of ad hoc expertise at common law in *R v Menzies* [1982] 1 NZLR 40, 49 (Cooke J for Cooke, McMullin, Somers JJ and Sir Clifford Richmond) and *Butera v DPP (Vic)* (1987) 164 CLR 180, 187–8 (Mason CJ, Brennan and Deane JJ). Consider its expansion under the *Uniform Evidence Law* (n 6) in *Eastman v The Queen* (1997) 76 FCR 9, 112–14 (Von Doussa, O'Loughlin and Cooper JJ); *R v Leung* (1999) 47 NSWLR 405, 412–15 [36]–[47] (Simpson J, Spigelman CJ agreeing at 406 [1], Sperling J agreeing at 418 [66]) ('*Leung*'); *Li* (n 31) 289 [57] (Ipp JA); *R v Madigan* [2005] NSWCCA 170, [92] (Wood CJ at CL); *R v Riscuta* [2003] NSWCCA 6, [34] (Heydon JA) ('*Riscuta*'); *R v Gao* [2003] NSWCCA 390, [23] (Greg James J).

[55] See Peter French and Helen Fraser, 'Why "Ad Hoc Experts" Should Not Provide Transcripts of Indistinct Forensic Audio, and a Proposal for a Better Approach' (2018) 42(5) *Criminal Law Journal* 298, 298–9. In most cases, these people were translators/interpreters working for police and investigative agencies or were themselves investigating officers: at 299–300.

time, permissive courts have allowed police and others engaged in the investigation to express their opinions about the identity of the speaker and, by analogy, the identity of persons in images.[56]

Ad hoc expertise is purported expertise.[57] Ad hoc experts are not known to be expert — ie significantly better than juries — at the task of identifying POIs in images.[58] In many, perhaps most, cases they do not appear conversant with (and do not reference) relevant literatures and research on image (or voice) comparison and associated dangers, or the most effective means of making expert opinions comprehensible.[59] They often undertake their comparisons in biasing conditions, such as where detectives ask them to compare a POI with a nominated suspect or suspects (as in *Tang*).[60] The concept of the ad hoc expert has been developed and applied for reasons of legal expediency — convenience to the courts, prosecutors and investigators — without attending to performance and the serious risks posed by the manner in which opinions are obtained, presented and evaluated. Ad hoc experts are prone to making (or attempting to make) categorical claims, such as claiming that an offender and alleged suspect are 'one and the same'.[61] Ignorant of — or inattentive to — research on face comparison and widespread criticism of the identification

[56] See, eg, *Leung* (n 54) 414–15 [46]–[47] (Simpson J, Spigelman CJ agreeing at 406 [1], Sperling J agreeing at 418 [66]); *Li* (n 31) 289 [57] (Ipp JA); *Murdoch v The Queen* (2007) 167 A Crim R 329, 356 [298] (Angel ACJ, Riley J and Olsson AJ) ('*Murdoch*'); *Tasmania v Chatters* [2013] TASSC 61, [64] (Porter J).

[57] Edmond and Martire, 'Knowing Experts' (n 40) 102.

[58] See Ruth Clutterbuck and Robert A Johnston, 'Exploring Levels of Face Familiarity by Using an Indirect Face-Matching Measure' (2002) 31(8) *Perception* 985, 990.

[59] See Gary Edmond, Kristy Martire and Mehera San Roque, 'Unsound Law: Issues with ("Expert") Voice Comparison Evidence' (2011) 35(1) *Melbourne University Law Review* 52, 66–7.

[60] Consider the suggestive circumstances in *Nasrallah v The Queen* [2015] NSWCCA 188, [6]–[8], [41] (McCallum J); *R v Korgbara* (2007) 71 NSWLR 187, 190 [8]–[10], 194 [23] (McColl JA); *Riscuta* (n 54) [23] (Heydon JA); *R v Drollett* [2005] NSWCCA 356, [63] (Simpson J); *R v Sterling* (2014) 19 DCLR(NSW) 74, 84–5 [34]–[38] (Yehia DCJ) ('*Sterling*'). Investigators and others may spend a good deal of time with the images and are almost always exposed to additional information about the suspect and the investigation: see, eg, at 84–5 [34]–[38]. In voice cases, they are often aware of the phone number and owner of the phone being intercepted, the use of names and nicknames in conversations, accents, as well as the beliefs of detectives: see, eg, *Riscuta* (n 54) [23] (Heydon JA). In addition, they are routinely provided with only one voice (or just the voices of those suspected) to compare: see, eg, at [4].

[61] *Tang* (n 9) 688 [28] (Spigelman CJ). There was other circumstantial evidence against Tang, including a fingerprint on stolen goods and admissions made by two of the offenders that implicated Tang: at 706 [98]–[99], 716 [157] (Spigelman CJ). The anatomist seems to have known about the evidence and investigation and was given images of the two offenders and Tang to compare with the POI: at 685 [16]–[17] (Spigelman CJ).

paradigm, ad hoc experts do not appear to appreciate the difficulty of the task, the level of error, or the magnitude of risks from cognitive bias.[62]

Not conspicuously engaged with relevant scientific research, Spigelman CJ concluded that the 'specialised knowledge' said to be underpinning the ad hoc expert's opinion was not able to support an identification.[63]

> Facial mapping, let alone body mapping, was not shown, on the evidence in the trial, to constitute 'specialised knowledge' of a character which can support an opinion of identity.[64]

Yet, that 'knowledge' was said to (somehow) enable the anatomist to identify and discriminate between anatomical features in the images.[65] This is a kind of admissibility compromise.[66] It allows a highly qualified witness, called by the prosecutor, and likely to be formally 'qualified' as an expert before the jury, to express speculative opinions insinuating identity.[67] Juries could hear testimony about purported similarities — such as the 'thick' lips and 'square' chin with 'a dimple' — and the absence of discernible differences, but the anatomist could not positively identify the POI as Tang.[68] The jury would receive no information, about the abilities of the witness or the frequency and independence of facial features, that might enable them to make sense of the opinions.[69]

[62] The identification paradigm is the widely criticised tendency of many traditional forensic sciences (eg fingerprint, document and ballistic comparison) to positively link a trace to a specific source without qualification: see Simon A Cole, 'Forensics without Uniqueness, Conclusions without Individualization: The New Epistemology of Forensic Identification' (2009) 8(3) *Law, Probability and Risk* 233, 236–7, 239–41.

[63] *Tang* (n 9) 713 [139]–[140] (Spigelman CJ).

[64] Tang (n 9) 714 [146] (Spigelman CJ). At 712 [135], Spigelman CJ acknowledges that '[t]here does appear to be a body of expertise based on facial identification'. Interestingly, there are no references in *Tang* (n 9) to the extensive scientific literature on face recognition and comparison (see above n 50) in supporting this claim and in facilitating the qualified admission of the anatomist's opinions.

[65] Ibid 709 [120].

[66] Simon A Cole, 'Splitting Hairs? Evaluating "Split Testimony" as an Approach to the Problem of Forensic Expert Evidence' (2011) 33(3) *Sydney Law Review* 459, 463–4.

[67] The anatomist was not called for the retrial: see Edmond et al, 'Law's Looking Glass' (n 1) 349 n 20. Tang was nonetheless convicted. That conviction presumably rested, at least in part, on the jury considering the images of the robbery with Tang sitting in the dock, but this time without the assistance of a purported expert.

[68] *Tang* (n 9) 687–8 [25] (Spigelman CJ).

[69] See ibid 685 [14], 687 [22], 704 [85], 709 [120] (Spigelman CJ). The trial judge and the NSWCCA agreed that the 'features displayed on the videotape or on the stills taken from the videotape were not such as the jury could themselves make a comparison': at 704 [85].

Judicial attempts to rescue identifications based on s 79(1) — by suggesting that the opinions of ad hoc experts are based on 'specialised knowledge'[70] — seem strained to say the least. At best, such attempts privilege training in an apparently related field and 'experience' with a particular person or set of images (or voice recordings).[71] This reliance is difficult to reconcile with s 76 of the *Uniform Evidence Law* and the limited exception s 79(1) provides for opinions based 'wholly or substantially' on 'specialised knowledge'.[72]

Tang is curious because endorsement of the common law idea of ad hoc expertise occurs alongside recourse to the United States ('US') Supreme Court's influential *Daubert v Merrell Dow Pharmaceuticals Inc* ('*Daubert*') jurisprudence.[73] *Daubert* was imported to assist with the definition of 'specialised knowledge':

> The word 'knowledge' connotes more than subjective belief or unsupported speculation. The term 'applies to any body of known facts or to any body of ideas inferred from such facts or accepted as truths on good grounds'.[74]

In *Daubert*, and the cases that followed it, the US Supreme Court explained that the word 'knowledge' (in r 702 of the *Federal Rules of Evidence*, 28 USC (1975)) imposed the need for reliability and validation of scientific evidence.[75] Remarkably, given this provenance, along with trends in comparable common law jurisdictions, the NSWCCA insisted that reference to 'specialised knowledge' in s 79 of the *Uniform Evidence Law* did not require trial judges or prosecutors to

[70] See ibid 709 [120], 712 [134]–[135] (Spigelman CJ).

[71] See Edmond and Martire, 'Knowing Experts' (n 40) 99.

[72] Judges sometimes suggest (or imply) that ad hoc experts have 'knowledge' of the images or the persons in the images, but really they only have exposure to, or experience with, the images, that anyone could obtain, and that is not based on identifiable knowledge: see, eg, *Tang* (n 9) 709 [120] (Spigelman CJ). To claim that experience with the images produces knowledge of a kind that could ground an opinion is alchemical — it transforms limited 'experience' into 'knowledge'. It renders the need for 'knowledge' in *Uniform Evidence Law* (n 6) s 79(1) redundant because experience becomes knowledge. To put this another way, these are opinions based on experience.

[73] 509 US 579 (1993) ('*Daubert*'). See Tang (n 9) 709 [120], 713–14 [139]–[146] (Spigelman CJ).

[74] *Tang* (n 9) 712 [138], quoting *Daubert* (n 73) 590 (Blackmun J for White, Blackmun, O'Connor, Scalia, Kennedy, Souter and Thomas JJ).

[75] In *Daubert* (n 73) 589–90 (Blackmun J for the Court), the US Supreme Court emphasised the need for validity and reliability and in *Kumho Tire Co Ltd v Carmichael*, 526 US 137 (1999) ('*Kumho*'), the Court confirmed that it was the word 'knowledge' (in r 702 of the *Federal Rules of Evidence*, 28 USC (1975)) that 'establishe[d] a standard of evidentiary reliability': *Kumho* (n 75) 147–9 (Breyer J for the Court). See also Gary Edmond, 'Specialised Knowledge, the Exclusionary Discretions and Reliability: Reassessing Incriminating Expert Opinion Evidence' (2008) 31(1) *University of New South Wales Law Journal* 1, 48–50.

consider the reliability of opinions.[76] Rather, '[t]he focus of attention must be on the words "specialised knowledge", not on the introduction of an extraneous idea such as "reliability"'.[77]

The upshot is that Australian prosecutors, trial judges and appellate courts are not concerned with reliability as a condition for the admission of opinion evidence in criminal proceedings.[78] Precisely what kind of knowledge, let alone specialised knowledge, is not reliable (or indexed to the tradition of justified (true) belief) is yet to receive elaboration.[79] Lack of attention to the validity and reliability of forensic science procedures, and the empirical basis of the words used to express opinions, form part of the unfortunate legacy of *Tang*.[80] Rather than require the proponent of the opinion (ie the purported expert) to provide evidence of actual expertise — here, empirical evidence of an ability to accurately compare faces and/or bodies to assist with the identification of persons in images — the Court in *Tang* instead relied upon proxies.[81] These proxies included the witness's training in anatomy, experience reconstructing faces from skulls,[82] the existence of general literature,[83] the fact that the witness had

[76] *Tang* (n 9) 712 [137] (Spigelman CJ). For a discussion, see Gary Edmond, 'Forensic Science Evidence, Adversarial Criminal Proceedings, and Mainstream Scientific "Advice"' in Darryl K Brown, Jenia Iontcheva Turner and Bettina Weisser (eds), *The Oxford Handbook of Criminal Process* (Oxford University Press, 2019) 761, 769.

[77] *Tang* (n 9) 712 [137] (Spigelman CJ).

[78] For further discussion, see Gary Edmond, 'The Admissibility of Forensic Science and Medicine Evidence under the *Uniform Evidence Law*' (2014) 38(3) *Criminal Law Journal* 136, 141–7; Gary Edmond, 'Regulating Forensic Science and Medicine Evidence at Trial: It's Time for a Wall, a Gate and Some Gatekeeping' (2020) 94(6) *Australian Law Journal* 427, 436 ('Regulating Forensic Science and Medicine Evidence').

[79] Cf Scott Brewer, 'Scientific Expert Testimony and Intellectual Due Process' (1998) 107(6) *Yale Law Journal* 1535, 1545–7; Tony Ward, 'Explaining and Trusting Expert Evidence: What Is a "Sufficiently Reliable Scientific Basis"?' (2020) 24(3) *International Journal of Evidence and Proof* 233, 233–4. See also Andrew Roberts, 'Probative Value, Reliability, and Rationality' in Andrew Roberts and Jeremy Gans (eds), *Critical Perspectives on the Uniform Evidence Law* (Federation Press, 2017) 63, 68–9 ('Probative Value').

[80] The approach to reliability in *Tang* (n 9) was endorsed in Victoria in *Tuite v The Queen* (2015) 49 VR 196, 217 [70] (Maxwell ACJ, Redlich and Weinberg JJA) ('*Tuite*') and reiterated in *Chen v The Queen* (2018) NSWLR 915, 926 [62] (Hoeben CJ at CL, Schmidt and Campbell JJ) — but was not substantively addressed in the High Court's *Honeysett* (n 8) judgment.

[81] See *Tang* (n 9) 709 [120] (Spigelman CJ).

[82] Reconstruction of faces from skulls seems to be a speculative and creative procedure: see Iscan (n 50) 68–9.

[83] *Tang* (n 9) 699–700 [70]–[72] (Spigelman CJ).

previously compared images for the police and other courts,[84] and her exposure to the images of the robbery for some unknown period of time.[85]

Tang was endorsed by common law courts in *Murdoch v The Queen* (in the Northern Territory) and *R v Dastagir* (in South Australia).[86] These decisions are not particularly attentive to common law admissibility rules that require a field of expertise and a qualified expert in that field who can assist the jury.[87] Judges in these jurisdictions also placed reliance on anatomical training and the admission of similar evidence in *Uniform Evidence Law* jurisdictions, notwithstanding differences in admissibility rules.[88] They were also willing to extend the scope of ad hoc expertise from the preparation of transcripts to the identification of POIs in video and sound recordings.[89] None focused on actual expertise in facial comparison for the purposes of identification. Of interest, in *Murdoch*, the Northern Territory Court of Criminal Appeal also endorsed the admissibility of opinions from persons who were very familiar with the defendant prior to the investigation ('familiars').[90] These familiars were allowed to identify the POI in poor-quality images at a truck stop as Murdoch.[91]

After *Tang*, the individuals (mostly anatomists) enlisted by investigators and called upon by prosecutors to testify did not always adhere to the judge-imposed restrictions limiting their opinions to descriptions of similar features.[92] It was common for anatomists (and others) to attribute significance to features

[84] Ibid 697–8 [61], [65].

[85] Ibid 709 [120].

[86] *Murdoch* (n 56) 352–5 [282]–[288] (Angel ACJ, Riley J and Olsson AJ); *Dastagir* (n 43) 94 [57] (Kourakis CJ, Vanstone and Stanley JJ).

[87] See, eg, *Clark v Ryan* (1960) 103 CLR 486, 491–2 (Dixon CJ); *R v Bonython* (1984) 38 SASR 45, 46–7 (King CJ) ('*Bonython*'). The formulation in *Bonython* (n 87) that refers to 'reliability' has not been effective in Australia or England and Wales: see, eg, *Tang* (n 9) 712 [137] (Spigelman CJ); Law Commission, *Expert Evidence in Criminal Proceedings in England and Wales* (Report No 325, 21 March 2011) 15 n 23 ('*Expert Evidence in Criminal Proceedings*').

[88] See *Dastagir* (n 43) 92–3 [49] (Kourakis CJ, Vanstone and Stanley JJ); *Murdoch* (n 56) 356 [298] (Angel ACJ, Riley J and Olsson AJ). South Australia has not adopted the *Uniform Evidence Law* (n 6), but the Northern Territory has: *Evidence (National Uniform Legislation) Act 2016* (NT).

[89] *Dastagir* (n 43) 91 [42]–[44], 92–3 [49] (Kourakis CJ, Vanstone and Stanley JJ); *Murdoch* (n 56) 356 [296]–[298].

[90] See *Murdoch* (n 56) 356 [351], 368–9 [367]–[369] (Angel ACJ, Riley J and Olsson AJ).

[91] Ibid 365 [348]–[351] (Angel ACJ, Riley J and Olsson AJ). See also *R v Marsh* [2005] NSWCCA 331, [31] (Studdert J).

[92] See, eg, *Murdoch* (n 56) 356 [300] (Angel ACJ, Riley J and Olsson AJ).

said to be similar — at least in their reports.[93] As those engaged in criminal activities began to wear disguises in response to the prevalence of security cameras (and facial mappers), the purported experts called by prosecutors began to express opinions that stimulated further judicial intervention.[94]

The judgment in *Morgan v The Queen* ('*Morgan*') is probably the clearest expression of judicial concern and reasoned exclusion.[95] The issue confronting the NSWCCA involved a different anatomist drawing attention to multiple similarities between a robber covered from head to toe in clothing and the Indigenous defendant — see Figure 3 below.[96] Professor Maciej Henneberg — the Wood Jones Chair of Anthropological and Comparative Anatomy at the University of Adelaide — listed the following features as shared between Morgan and the POI in his report:

> Person of interest is an adult male of heavy body build. His shoulders and hips are wide. He has a prominent abdomen but his upper and lower limbs, especially in their distal segments, are not thick. This suggests centripetal pattern of body fat distribution. This pattern consists of the deposition of most body fat on the trunk while limbs remain relatively thin. His head and face were covered by a garment well adhering to the surface of the skin. This enabled me to make observations of the head shape, nose and face profile. His head is dolichocephalic (elongated) in the horizontal plane (viewed from above). His nose is wide and rather prominent while his face has straight profile (orthognathic). He is right-handed in his actions and carries himself straight.[97]

In addition, Professor Henneberg's testimony stated that 'Australian Aboriginals' have

> long thin limbs and that, when they put on weight due to a Western diet, it tended to be concentrated in the trunk, the process he described as centripetal fat distribution. He observed these features in the offender in the CCTV footage and,

[93] Consider, for example, the expert reports prepared for *Morgan* (n 22) 45 [76], 47–50 [83]–[96] (Hidden J) and *Honeysett v The Queen* (2013) 233 A Crim R 152, 156 [19]–[21] (Macfarlan JA) ('*Honeysett (NSWCCA)*').

[94] *Morgan* (n 22) 45 [74], 60–1 [143]–[146] (Hidden J).

[95] Ibid 60–1 [143]–[146].

[96] See ibid 45 [74]–[76] (Hidden J). The NSW Police and Office of the Director of Public Prosecutions appear to have strategically 'dropped' Dr Sutisno after the implicit criticism in *Tang* (n 9) 712 [136]–[137] (Spigelman CJ).

[97] *Morgan* (n 22) 45 [74] (Hidden J).

> later, in the images of the appellant. It is apparent from the [reference] photos of the appellant in evidence that he is Aboriginal.[98]

In his report and testimony, Professor Henneberg attributed significance, namely 'a high degree of anatomical similarity', to the features he described.[99]

Figure 3: Robber alleged to be Morgan (centre)

The *Morgan* Court was not especially interested in the validity and reliability of the procedure or evidence of Professor Henneberg's abilities.[100] Rather than require analysis of the method (or 'approach'), the Court drew attention to the need to assist the jury (as discussed in *Smith*)[101] and the lack of 'satisfactory' explanation,[102] remarking that

98 Ibid 46 [80].

99 Ibid 46 [79]. These are not especially helpful terms. They bring to mind terms like 'to a reasonable degree of medical certainty' and 'to a reasonable degree of ballistic certainty' that were scathingly criticised by scientists reviewing the reporting of opinions by expert witnesses in the US: see, eg, President's Council of Advisors on Science and Technology, Executive Office of the President, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (Report to the President, September 2016) 19, 30–1, 54–5, 86, 145 ('*PCAST Report*').

100 *Morgan* (n 22) 60–1 [144]–[145] (Hidden J).

101 *Smith* (n 20) 663–4 [41] (Kirby J).

102 *Morgan* (n 22) 60–1 [144] (Hidden J).

> this is not the occasion to examine the science of body mapping or to undertake some general appraisal of Professor Henneberg's approach. … [T]he question is whether he had specialised knowledge, *beyond the reach of lay people*, which he brought to bear in arriving at his opinion. That question fell to be answered by reference to the task which he undertook.
>
> That task was *to make an anatomical comparison between relatively poor quality CCTV images of a person covered by clothing from head to foot with images of the appellant*. Applying his specialised knowledge, Professor Henneberg claimed, he was able to detect not just a measure of similarity but 'a high level of anatomical similarity' between the two persons. How he was able to do that when no part of the body of the offender in the CCTV images was exposed was, in my view, never satisfactorily explained.[103]

Why an appeal considering the admissibility of opinions proffered by a face and body mapper was not 'the occasion to examine the science of body mapping',[104] or the general methods, might be thought to raise a question. How can a court determine if an opinion is based on 'specialised knowledge' without considering the 'approach' and its correspondence with knowledge?[105]

In *Morgan*, the NSWCCA was nevertheless attentive to the precise 'task' that the expert performed — in italics in the extract above — though the contention that the opinion was not *based* on 'specialised knowledge' appears declaratory:

> Whatever might be made of the professor's observations of the offender's body shape through his clothing, his observations about the shape of his head and face were clearly vital to his conclusion that there was a high degree of anatomical similarity between that person and the appellant. It does not appear to me that those observations could be said to be based upon his specialised knowledge of anatomy.[106]

The fact that the head and body were covered compounded the issues. The professor's experience with body shapes and sizing for the clothing industry did not redeem his interpretations:

> Professor Henneberg's evidence about his experience of the clothing industry … appears to be confined to the size and hang of garments, and their relation to 'body shape and posture'. … [T]he evidence does not convey that his experience

[103] Ibid 60 [139]–[140] (Hidden J) (emphasis added).

[104] Ibid 60 [139].

[105] On the requirements of the form in which an opinion is presented, see *HG v The Queen* (1999) 197 CLR 414, 427 [39] (Gleeson CJ).

[106] *Morgan* (n 22) 60–1 [144] (Hidden J).

> extends to the observation of anatomical features of the head and face of a person whose head is entirely covered by a garment such as a balaclava.[107]

Professor Henneberg's opinions were deemed inadmissible.[108]

Shortly thereafter, a differently configured NSWCCA found Professor Henneberg's bare similarity evidence, in another conviction involving a well-disguised armed robber, to be admissible.[109] The NSWCCA's *Honeysett v The Queen* decision is difficult to reconcile with *Morgan*. The High Court granted special leave.[110] A unanimous High Court deemed the opinions of Professor Henneberg inadmissible because *they were said* not to be 'based wholly or substantially on his specialised knowledge within s 79(1) of the *Uniform Evidence Law*. It was an error of law to admit the evidence'.[111]

Figure 4: Robber alleged to be Honeysett (top left)

[107] Ibid 60 [143].

[108] Ibid 61 [145]–[146].

[109] *Honeysett (NSWCCA)* (n 93) 164 [66]–[68] (Macfarlan JA, Campbell J agreeing at 165 [77], Barr AJ agreeing at 165 [78]).

[110] *Honeysett* (n 8) 126 [3] (French CJ, Kiefel, Bell, Gageler and Keane JJ). Leave was not granted in the South Australian case of *Dastagir* (n 43): see Transcript of Proceedings, *Dastagir v The Queen* [2014] HCATrans 130, 385–400 (French CJ).

[111] *Honeysett* (n 8) 138 [46] (French CJ, Kiefel, Bell, Gageler and Keane JJ).

On comparing the crime scene images (see Figure 4) with reference images of Honeysett, the professor purported to identify several similar features, including: somatotype (ectomorphic), height, posture (lumbar lordosis), hair length, skull shape (dolichocephalic), handedness and skin colour (once again the defendant was an Indigenous Australian).[112] The Court's not entirely helpful explanation of why the comparison and opinions about alleged similarities are not admissible is extracted in full below:

> Professor Henneberg's opinion was not based on his undoubted knowledge of anatomy. Professor Henneberg's knowledge as an anatomist, that the human population includes individuals who have oval shaped heads and individuals who have round shaped heads (when viewed from above), did not form the basis of his conclusion that Offender One and the appellant each have oval shaped heads. That conclusion was based on Professor Henneberg's subjective impression of what he saw when he looked at the images. This observation applies to the evidence of each of the characteristics of which Professor Henneberg gave evidence.[113]

This extract exposes a court struggling to provide meaningful explanation for exclusion. All human-based image comparisons will require 'subjective impressions' of what is seen.[114] While it is true that anatomical training, study, and experience and even anatomical knowledge may not be of value when engaging in comparisons, their actual value is uncertain — ie unknown. In the terminology from *Smith*, we do not know if Professor Henneberg's opinions are relevant.[115] Can he actually do it (better than the jury)? In our terms, the expertise is *purported*. We are not told about 'the basis' of the opinion, the relevant 'knowledge', or Professor Henneberg's actual ability at image comparison for the purpose of identification.[116] While we agree that his opinion(s) should be excluded — for us, because it is purported expertise — the implications of the

[112] Ibid 127 [9], 129 [15] (French CJ, Kiefel, Bell, Gageler and Keane JJ). The accused's skin colour was an obvious feature of the reference images provided to the anatomist.

[113] Ibid 138 [43] (French CJ, Kiefel, Bell, Gageler and Keane JJ).

[114] Ibid.

[115] See *Smith* (n 20) 655–6 [10]–[12] (Gleeson CJ, Gaudron, Gummow and Hayne JJ). As explained in *Volpe v The Queen* [2020] VSCA 628 ('*Volpe*'), the fact-finder must be careful about accepting opinions (and infusing them with probative value): at [70]–[74] (Priest, T Forrest and Weinberg JJA).

[116] There is very limited engagement with the basis for the opinion: *Honeysett* (n 8) 129–30 [15]–[19] (French CJ, Kiefel, Bell, Gageler and Keane JJ). For a more detailed analysis of the importance of identifying the basis of an opinion, see *Makita (Australia) Pty Ltd v Sprowles* (2001) 52 NSWLR 705, 730–3 [60]–[69] (Heydon JA); *Davie v Magistrates of Edinburgh* 1953 SC 34, 40 (Lord President Cooper).

Court's reasoning are unclear. Moreover, we cannot be confident that the decision extends beyond anatomists (or this anatomist) or applies in cases where the defendant is not well disguised.[117]

The High Court's response, like the decisions in *Tang* and *Morgan*, is declaratory. Without determining whether the anatomist possesses expertise in comparing persons in images, it simply declares that the opinion is not based on 'specialised knowledge'.[118] The judgment provides no meaningful guidelines and tells us nothing about what kind of 'specialised knowledge' might ground an admissible opinion about identity.[119] The judgment does not engage with similarity constraints or the relationship between knowledge and reliability (which were left open in *Tang*). Legally, it is unclear what future prosecutors, defence counsel and judges should look for when trying to determine if an opinion is based on 'specialised knowledge' (and whether that 'knowledge' is based on 'training, study or experience').[120] Like the Court in *Morgan*, the High Court in *Honeysett* disavowed the need to deal in a principled manner with either the scope of 'specialised knowledge',[121] or to address whether there was a place for ad hoc expertise within the framework offered by s 79 of the *Uniform Evidence Law*.[122]

[117] Indeed, a different court might have made more of the professor's comparative anthropological research or experience with cameras.

[118] *Honeysett* (n 8) 138 [43] (French CJ, Kiefel, Bell, Gageler and Keane JJ). Ironically, in *Tang* (n 9) the jury asked questions along these very lines, but the appropriate concessions, disclosure and explanation were not forthcoming. The jury asked: 'Accepting Dr Sutisno's qualifications should we therefore accept her methodology?': at 695 [50] (Spigelman CJ). At 701 [74], the jury also asked:

> Was there any photo anthropometry performed in comparing the surveillance images and forensic photos of the accused, what were the results? How accurate is morphology analysis as a technique? What percentage of cases are correct matches of persons versus incorrect matches? Could we please ask Dr Sutisno how many matching morphological features she needs to form the opinion that two photos are the same person, what would be the minimum?

[119] Cf *Federal Rules of Evidence* (n 75) r 702, following revision in 2000, or the criteria advanced in *Daubert* (n 73) 592–4 (Blackmun J for White, Blackmun, O'Connor, Scalia, Kennedy, Souter and Thomas JJ), namely: testable and tested; peer-reviewed and published; a known error rate; the existence of standards; and scope to consider whether the procedure was generally accepted.

[120] The contention that the professor's opinion was admissible as ad hoc expertise was strategically abandoned by the Crown on the basis of his limited exposure to the images. In consequence, the High Court did not consider the issue: *Honeysett* (n 8) 138–9 [47]–[48] (French CJ, Kiefel, Bell, Gageler and Keane JJ). Interestingly, the High Court noted that in the NSWCCA 'Macfarlan JA said that Professor Henneberg's detailed examination of the CCTV footage over a lengthy period had qualified him as such': at 138–9 [47].

[121] Ibid 131–2 [23]–[24] (French CJ, Kiefel, Bell, Gageler and Keane JJ).

[122] Ibid 138–9 [47]–[48].

B *Sections 78 and 137*

Compounding problems, courts in *Uniform Evidence Law* jurisdictions have begun to rationalise the admission of investigators' opinions — most conspicuously in relation to the identification of persons captured speaking on telephone and other voice intercepts — via the exception for lay opinions: namely, s 78.[123] This approach overlaps with the idea of the ad hoc expert, and enables persons without any 'specialised knowledge' or 'training, study or experience' to present their opinions about identity (and even the words allegedly spoken and sometimes their meaning) on the basis of listening or repeated listening to a recording.[124] This development is problematic because it has enabled investigators, and interpreters working with them, to express their opinions about the identity of the speaker.[125] The individuals proffering their opinions via s 78 are not *directly* perceiving the matter or event. No evidence is presented or required to support expertise in voice comparison.[126] They are not required to satisfy the *Code of Conduct for Expert Witnesses*[127] and, like the ad hoc experts speaking about a POI in images (via s 79), tend to have little if any idea about the difficulty of the task, relevant scientific research, or notorious risks known to

[123] See, eg, *Tran v The Queen* [2016] VSCA 79, [102]–[104] (Weinberg, Santamaria and McLeish JJA) ('*Tran*'). At common law, lay opinion was limited to direct or sensory witnesses: see *Evidence Interim Report* (n 35) 410 [739]. The current interpretation of s 78 of the *Uniform Evidence Law* (n 6), applied to recordings, is inconsistent with this tradition and does not seem to follow from the text, context or purpose of the section. It appears to enable *anyone* who watches or listens to a recording — more than a jury is likely to — to proffer an opinion.

[124] *Uniform Evidence Law* (n 6) s 79. Those who are allowed to express their opinions via s 78 are sometimes described as ad hoc experts: see *Tran* (n 123) [102]–[104], [111] (Weinberg, Santamaria and McLeish JJA). Consequently, ad hoc experts' evidence might be admitted via s 78 or s 79(1). On the interpretation of meaning through repeated listening, see *Dodds v The Queen* (2009) 194 A Crim R 408, 412–14 [19]–[26] (McClellan CJ at CL); *Keller v The Queen* [2006] NSWCCA 204, [23]–[24] (Studdert J). The NSWCCA's view on the opinions of dog-handlers about the meaning of their animals' behaviour differed in *Muldoon v The Queen* (2008) 192 A Crim R 105, 117–18 [38]–[41] (Hodgson JA) and *R v Benecke* (1999) 106 A Crim R 282, 284 [22] (Barr J).

[125] *Tran* (n 123) [79], [83]–[86], [104]–[105] (Weinberg, Santamaria and McLeish JJA); *Nguyen* (n 32) 415 [40] (Basten JA), 425 [101]–[106] (Schmidt J); *Kheir v The Queen* (2014) 43 VR 308, 324 [67]–[71] (Maxwell P, Redlich and Beach JJA) ('*Kheir*'). See also *Riscuta* (n 54) [26]–[27], [41], [60]–[61] (Heydon JA); *R v El-Kheir* [2004] NSWCCA 461, [94]–[97], [111]–[114] (Tobias JA); *Bulejcik v The Queen* (1996) 185 CLR 375, 382–3 (Brennan CJ).

[126] Where those proffering opinions are translators (who understand the language on the recording), they might have an advantage over the jury in voice comparison, but studies confirm that even such comparisons are surprisingly error-prone, especially when undertaken in the suggestive conditions of investigations. The probative value of such opinions will often be low and the risks of error and unfair prejudice high. Voice comparison seems to be even more error-prone than identification by images: see generally Edmond and San Roque (n 51) 11–13.

[127] Federal Court of Australia, *Expert Evidence Practice Note*, 25 October 2016, Annexure A.

specialists.[128] Section 78 does not require knowledge.[129] Revealingly, s 78 witnesses have demonstrated a striking tendency to categorically identify speakers and to maintain high levels of confidence (bearing limited correspondence with studies of general abilities) when cross-examined,[130] for they are not restricted to describing similarities.[131]

The opinions deemed inadmissible by virtue of s 79(1) in *Honeysett* and *Morgan* would appear to be admissible by virtue of s 78. By analogy with the interpretation of a voice recording, the anatomist is merely describing what they see, hear or otherwise perceive about a recording of the 'matter or event' (though, here a photograph or video) and in the same way that it is said to be 'necessary' to receive the opinions of investigators about who is speaking, it would seem to be necessary to receive the opinion of the anatomist about the identity of the POI in the image(s).[132] Ironically, the constraints imposed in *Tang*, limiting the opinions to similarities and differences, do not apply to opinions admitted according to the exception for lay opinions provided by s 78 and its common law equivalents.[133] Interpreted in this way, s 78 threatens to undermine the requirements imposed by s 79(1). It enables investigators (and others) to testify without having to possess or attend to 'specialised knowledge'.[134] This sidesteps Kirby J's reasoning in *Smith* as well as subsequent authority from *Lithgow City Council v Jackson* focused on 'necessity' and directness, and constitutes a striking inconsistency.[135] It allows those with little (if any) knowledge, but with purported abilities, to express their opinions in the strongest possible terms.[136]

[128] *Kheir* (n 125) 316–17 [39]–[48] (Maxwell P, Redlich and Beach JJA). See also Geoffrey Stewart Morrison and William C Thompson, 'Assessing the Admissibility of a New Generation of Forensic Voice Comparison Testimony' (2017) 18(2) *Columbia Science and Technology Law Review* 326, 341–7; Claudia Rosas, Jorge Sommerhoff and Geoffrey Stewart Morrison, 'A Method for Calculating the Strength of Evidence Associated with an Earwitness's Claimed Recognition of a Familiar Speaker' (2019) 59(6) *Science & Justice* 585, 585.

[129] Those who provide interpretations based on recordings may have more limited perspectives than those who were present and able to perceive the events as they unfolded.

[130] See Edmond, 'Regulating Forensic Science and Medicine Evidence' (n 78) 433.

[131] This is why their opinions are said to be necessary. See the discussion of s 78(b) in *R v Whyte* [2006] NSWCCA 75, [36]–[37] (Spigelman CJ), [56]–[57] (Simpson J) ('*Whyte*').

[132] While it might be possible to distinguish voice from image comparisons, such distinctions would tend to be formalistic and artificial. They would not assist legal practitioners and investigators.

[133] See *Whyte* (n 131) [35]–[38] (Spigelman CJ).

[134] *Uniform Evidence Law* (n 6) s 79(1).

[135] *Lithgow* (n 35) 370–2 [45]–[46], 374–5 [50]–[54] (French CJ, Heydon and Bell JJ).

[136] Edmond, 'Regulating Forensic Science and Medicine Evidence' (n 78) 433.

Ad hoc expertise tends to be represented as an obvious — even common sense — category.[137] It is, however, incompatible with the express terms of ss 76 and 78–9, as well as their context and purpose. Section 76 covers the field.[138] Section 78 requires the observation to be direct — or it should.[139] Section 79 requires 'specialised knowledge' as well as 'training, study or experience'. Recourse to ad hoc expertise leads to the admission of opinions based on nothing more than repeated exposure represented as either 'knowledge' or 'experience' (when admitted via s 79) and 'necessary' (when admitted via s 78). Ad hoc experts, whether expressing opinions under s 78 or s 79(1) are typically unfamiliar with the accuracy of their opinions, relevant scientific research, or dangers of cognitive bias.[140] They are purported experts. They do not know the value of their opinions.[141]

Problems are compounded by the reluctance to exclude opinions about POIs in images following objections based on s 137 (and s 135). Section 137 requires a trial judge to exclude evidence (of any kind) where its probative value is outweighed by the danger of unfair prejudice to the defendant.[142] Authority, specifically *IMM v The Queen* ('*IMM*'), prevents trial judges from considering the reliability of the evidence or the credibility of the witness when undertaking

[137] See Edmond and San Roque (n 51) 33.

[138] '[T]he duty of a court is to give the words of a statutory provision the meaning that the legislature is taken to have intended them to have': *Project Blue Sky Inc v Australian Broadcasting Authority* (1998) 194 CLR 355, 384 [78] (McHugh, Gummow, Kirby and Hayne JJ). Part 3.3 of the *Uniform Evidence Law* (n 6) is concerned with opinion evidence and s 76(1) appears to cover the field for opinions adduced 'to prove the existence of a fact about the existence of which the opinion was expressed'.

[139] In *Smith* (n 20), Kirby J deemed the opinions of the police officers inadmissible on the basis that the possible exception for opinions based on what a witness 'saw, heard or otherwise perceived about a matter or event' was not satisfied: at 669–70 [59]–[61], quoting *Evidence Act 1995* (NSW) s 78(a). Intermediate courts of appeal, for example in *Kheir* (n 125) 323 [65] (Maxwell P, Redlich and Beach JJA), appear to have overlooked *Lithgow* (n 35), in which the High Court indicates that s 78(a) applies only to direct sensory witnesses: at 368 [41] (French CJ, Heydon and Bell JJ).

[140] See, eg, Edmond and San Roque (n 51) 31. It is possible that some types of exposure may actually enhance performance, but courts have not engaged with relevant research or required any insight about this. Rather, repeated exposure is assumed to improve performance: see, eg, *Tang* (n 9) 709 [120] (Spigelman CJ). Risks from suggestive processes and error are ignored or left as issues for the defence and the trier of fact.

[141] Cross-examination and judicial directions are unlikely, and perhaps incapable, of rectifying this situation: see Edmond and Wortley (n 12) 516–17.

[142] It requires an objection, though defence counsel have not been particularly effective in their arguments on probative value or unfair prejudice flowing from the interpretation of images (and sounds). An instructive exception is *Pentland v The Queen* [2020] QSCPR 10, [26], [78] (Lyons SJA).

the balancing exercise.[143] This controversial approach applies to forensic science and medicine evidence.[144] Without insight into the value of an opinion or the actual risk of error, trial judges are required to determine the capacity of such evidence — in order to take it 'at its highest'.[145] Lack of familiarity with scientific research on the difficulty of identifying strangers has enabled judges to routinely find the probative value of opinions to outweigh (unknown) dangers to the defendant.[146] Judges do not typically consider the opinions of purported experts as 'weak' or risky.[147] They have tended to admit opinions on the basis that any weaknesses or limitations will be exposed (and conveyed) by trial safeguards — such as cross-examination and judicial warnings — notwithstanding the fact that these are rarely informed by mainstream scientific knowledge and almost never expose, for consideration, real risks of error and misunderstanding.[148]

Based on the published decisions, no Australian court has ever required or been presented with information on whether those interpreting images can accurately describe features or identify specific persons.[149] No court has discussed

[143] (2016) 257 CLR 300, 313 [44], 314–15 [48]–[50] (French CJ, Kiefel, Bell and Keane JJ) ('*IMM*'). Ironically, the failure of many of the proffered image witnesses to address validity and reliability issues is tightly coupled with their credibility. One might reasonably wonder about the credibility of a (purported) expert who does not recognise or draw attention to validity and proficiency.

[144] See Gary Edmond, 'Icarus and the *Evidence Act*: Section 137, Probative Value and Taking Forensic Science Evidence "at Its Highest"' (2017) 41(1) *Melbourne University Law Review* 106, 111; Roberts, 'Probative Value' (n 79) 75; David Hamer, 'The Unstable Province of Jury Fact-Finding: Evidence Exclusion, Probative Value and Judicial Restraint after *IMM v The Queen*' (2017) 41(2) *Melbourne University Law Review* 689, 697–9.

[145] *IMM* (n 143) 313 [44] (French CJ, Kiefel, Bell and Keane JJ).

[146] See, eg, *Leung* (n 54) 415 [47] (Simpson J). Similar mistaken assumptions apply to voice comparison and the production of transcripts (ie what was spoken) on voice recordings: see generally Edmond and San Roque (n 51) 10–13.

[147] See, eg, *Li* (n 31) 291 [71]–[72] (Ipp JA). There may be limited scope to do so (provided it is not characterised as a reliability assessment): see *IMM* (n 143) 314–15 [50] (French CJ, Kiefel, Bell and Keane JJ); *R v XY* (2013) 84 NSWLR 363, 376–7 [48] (Basten JA).

[148] See *Li* (n 31) 291 [71]–[72]. On the limits of cross-examination, see Dawn McQuiston-Surrett and Michael J Saks, 'The Testimony of Forensic Identification Science: What Expert Witnesses Say and What Factfinders Hear' (2009) 33(5) *Law and Human Behavior* 436, 439; *Expert Evidence in Criminal Proceedings* (n 87) 5 [1.20], 6 [1.24]. See also Committee on Identifying the Needs of the Forensic Science Community, National Research Council, *Strengthening Forensic Science in the United States: A Path Forward* (Report, 2009) 53, 85–110 ('*NRC Report*'); Gary Edmond et al, 'Forensic Science and the Limits of Cross-Examination' (2019) 42(3) *Melbourne University Law Review* 858, 869.

[149] Following *Aytugrul v The Queen* (2012) 247 CLR 170 ('*Aytugrul*'), appellate courts are circumscribed by what the parties and their witnesses bring into proceedings during the

scientific research, (lack of) error rates and related risks associated with unfamiliar face comparison. We do not know whether those allowed to, or those prevented from, proffering opinions have relevant abilities. Instead of focusing on reliability, Australian judges have been distracted by other factors — eg qualifications, experience (including legal experience), and repeated exposure in suggestive conditions — that are secondary to the question of whether the witness has a heightened ability.[150] Thus, Australian courts are curiously inattentive to the only criteria that actually matter: *are these individuals really experts at image comparison, how good are they, and where is the evidence (ie 'knowledge') that supports the claimed ability?*[151]

III Unfamiliar Face Matching: Relevant Scientific Research

Courts can draw on three main sources of scientific research to inform decisions relating to face comparison evidence. First, a large body of scientific research conducted since the early 1990s has examined the accuracy of ordinary (or lay) persons at facial comparison (or 'face matching' as it is often described by research scientists) and the factors that affect their accuracy.[152] Secondly, in more recent years studies have examined the accuracy of various groups involved professionally in unfamiliar face matching.[153] These groups range from

trial: at 184 [22] (French CJ, Hayne, Crennan and Bell JJ). See also Gary Edmond, David Hamer and Emma Cunliffe, 'A Little Ignorance Is a Dangerous Thing: Engaging with Exogenous Knowledge Not Adduced by the Parties' (2016) 25(3) *Griffith Law Review* 383, 404.

150 See, eg, *Tang* (n 9) 709 [120] (Spigelman CJ). See also *PCAST Report* (n 99) 6, for discussion of a similar experience in the US.

151 Martire and Edmond, 'Rethinking Expert Opinion Evidence' (n 41) 984–7. Focus is occasionally directed towards the specific task, as in *Morgan* (n 22) 60 [139]–[141] (Hidden J), though rarely on evidence of actual ability. See also the discussion of knife wounds in *Gilham v The Queen* [2012] NSWCCA 131, [330]–[345], [350] (McClellan CJ at CL, Fullerton and Garling JJ), where the Court was similarly focused on the specific task (ie identifying similarity of wounds).

152 Lay persons, sometimes described as 'novices', are frequently tertiary students engaged by experimenters: see, eg, David White et al, 'Perceptual Expertise in Forensic Facial Image Comparison' (2015) 282(1814) *Proceedings of the Royal Society B* 20151292:1–8, 2. For a review, see Peter JB Hancock, Vicki Bruce and A Mike Burton, 'Recognition of Unfamiliar Faces' (2000) 4(9) *Trends in Cognitive Sciences* 330, 330–5.

153 For examples of this experimental work, see P Jonathon Phillips et al, 'Face Recognition Accuracy of Forensic Examiners, Superrecognizers, and Face Recognition Algorithms' (2018) 115(24) *Proceedings of the National Academy of Sciences* 6171 ('Face Recognition Accuracy'); Alice Towler, David White and Richard I Kemp, 'Evaluating the Feature Comparison Strategy for Forensic Face Identification' (2017) 23(1) *Journal of Experimental Psychology* 47 ('Evaluating the Feature Comparison Strategy'); David White et al, 'Passport Officers' Errors in Face

staff performing image comparison incidentally in their daily work (eg nightclub bouncers) to highly trained and experienced specialist facial examiners whose work focuses primarily (and sometimes exclusively) on interpreting images, writing formal reports and (outside Australia) testifying in criminal proceedings.[154] Thirdly, since the 1990s, the accuracy of automated facial recognition systems (ie algorithms) has been rigorously evaluated.[155]

The technicalities of this research sometimes make it difficult to extract the data and findings most pertinent to legal practice. So, in this section, we have provided an overview focusing on the relative accuracy of the different types of individuals and systems that might be called upon to assist with the identification of POIs in images in investigations and criminal prosecutions. Some of the findings we describe are well established — that is, replicated across many studies — but some areas of research are in their infancy. We have endeavoured to address the relative scientific certainty of the various research findings.

A *Factors That Affect Accuracy in Face Comparison*

Initially it is important to recognise that the accuracy of all sources of face comparison evidence is affected by a variety of factors. Any assessment of identification evidence requires careful consideration of the following sources of variation: (i) how familiar the decision-maker is with the person to be identified; (ii) the quality and quantity of the images relied upon; and (iii) interactions between the demographics of persons in the images and the decision-maker.[156] In the following pages we consider the levels of accuracy that can be expected from different types of image comparison before reviewing the relative accuracy of the different groups and systems.

Matching' (2014) 9(8) *PLoS ONE* e103510:1–6; White et al, 'Perceptual Expertise in Forensic Facial Image Comparison' (n 152). For a review of the literature on professional face matchers, see David White, Alice Towler and Richard I Kemp, 'Understanding Professional Expertise in Unfamiliar Face Matching' in Markus Bindemann (ed), *Forensic Face Matching: Research and Practice* (Oxford University Press, 2021) 62.

154 See, eg, Phillips et al, 'Face Recognition Accuracy' (n 153) 6171–2.

155 These results are reported in computer science and engineering journals and annual reports published by the United States National Institute of Standards and Technology ('NIST'). Copies of the NIST reports can be found at 'Face Recognition Vendor Test (FRVT)', *National Institute of Standards and Technology* (Web Page, 30 November 2020) <http://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt>, archived at <https://perma.cc/G2U3-B4EQ>.

156 That is, believed characteristics of persons in the images.

1 *Familiarity with the Face*

A very important moderator of accuracy in face identification tasks is how familiar a viewer is with the particular face.[157] Humans are generally very good at identifying the faces of individuals who are known or *familiar* to them.[158] Deciding whether the two images on the left of Figure 5 ('pair A') portray the same person will not be particularly challenging for most people reading this article, notwithstanding substantial differences in age, pose, expression, image quality, make-up and distance from the camera.[159] On the other hand, most people find a similar question, about the faces on the right ('pair B') to be very difficult.[160] Yet, the two images on the right were taken just minutes apart, under the same lighting conditions, with the same subject-to-camera distance and in similar pose.[161] Around half of those asked (incorrectly) report that these are images of different people.[162]

This example highlights difficulties with the kinds of unfamiliar face matching decisions that are routinely made in courts. It also demonstrates the transformative effect that being familiar with a face has on face identification accuracy. Research shows that the level of familiarity is correlated with accuracy.[163] In studies that have directly compared the identification of familiar and unfamiliar faces by asking subjects to compare still images with CCTV images, performance with personally familiar faces is substantially better (with accuracy around 90%) than with unfamiliar faces (around 70% correct).[164]

157 Hancock, Bruce and Burton (n 152) 330.

158 Ibid.

159 That will, of course, depend on their age and where they live. If this article were read by young people or in 50 years from now, the first comparison might be difficult, perhaps even more so than the other.

160 White, Towler and Kemp (n 153) 62–3, citing A Mike Burton, David White and Allan McNeill, 'The Glasgow Face Matching Test' (2010) 42(1) *Behavior Research Methods* 286, 287–8.

161 Burton, White and McNeill (n 160) 287.

162 Ibid. Experimenters know the correct answer (ie the 'ground truth') about whether these are the same or different persons and so can calculate accuracy.

163 See, eg, A Mike Burton et al, 'Face Recognition in Poor-Quality Video: Evidence from Security Surveillance' (1999) 10(3) *Psychological Science* 243, 244–5, 247.

164 Vicki Bruce et al, 'Matching Identities of Familiar and Unfamiliar Faces Caught on CCTV Images' (2001) 7(3) *Journal of Experimental Psychology* 207, 212. These values for accuracy depend on image quality: at 208, 216–17. It is the differences in image quality that are of most interest in the results: see also Burton et al (n 163) 244–6.

Figure 5: Two face matching decisions — do these pairs of images show the same person or different people? The images in pair A illustrate how familiarity can simplify the task. Deciding whether the images in pair B depict the same person (ie 'match') is the kind of task presented in standard face comparison studies.

The level of familiarity with the suspect/defendant is therefore an important predictor of accuracy in face identification by those endeavouring to determine the identity of the POI in images.[165] Significantly, familiars and most purported experts have different types and levels of exposure to the faces they might be asked to identify. Their 'familiarity' is usually obtained in quite different circumstances. Familiars are exposed to a face (and a body) across a wide variety of interactions and settings, often over months or years. Think of a school friend, sibling or partner. Purported experts, on the other hand, usually have more limited exposure obtained in less varied conditions. Their exposure might be based on watching videos over radically foreshortened time periods — eg a police interview, the execution of a search warrant or a bank robbery. They may even be limited to studying just a single photo (or frame).[166] A number of studies have shown that identification accuracy is poorer for people who are moderately familiar compared with those who are very familiar.[167] Performance improves from near chance in poor-quality CCTV to high levels of accuracy when the questioned face is familiar.[168] Importantly, familiarity obtained by studying images of unfamiliar faces — akin to the experience of ad hoc experts and

[165] Recall the acquaintances in *Murdoch* (n 56) 366 [353]–[357] (Angel ACJ, Riley J and Olsson AJ).

[166] In the English case of *R v Atkins* [2010] 1 Cr App R 117, the medical artist called to identify the POI had access to a single, low-quality frame, taken at night, for comparison: at 120–1 [5]–[7] (Hughes LJ). For a closer analysis of this case, see the discussion in Edmond et al, '*Atkins v The Emperor*' (n 1) 150–6. In other cases, as in the example of the expert's photo board from Figure 2, we should be attentive to the reasons particular frames were selected for comparison.

[167] Clutterbuck and Johnston (n 58) 990.

[168] Burton et al (n 163) 247. See also Bruce et al (n 164) 210–11.

others involved in the investigation (ie investigative familiars) — provides relatively meagre benefits.[169] Ad hoc experts are not genuine familiars.[170]

In addition to the amount and kind of exposure it is also important to consider the length of time that has passed since a familiar last encountered the person in question.[171] Familiarity with faces fades with lack of exposure.[172]

2 *Image Quality and Quantity*

Another key factor influencing face matching accuracy is image quality. One would hope that this is already taken into account by courts, as common sense suggests that identification decisions from poor-quality images are unlikely to be reliable. The cases we reviewed in Part II suggest that this limitation has not received appropriate consideration. Australian courts admit low-quality images (eg in *Tang*) and images of well-disguised individuals (eg in *Morgan* and *Honeysett*), leaving the central issue of identity to the jury.[173]

Decades of research on the accuracy of unfamiliar face matching confirms that performance is dramatically reduced by low-quality images.[174] When identifying people from images and videos, accuracy is moderated by environmental and image capture conditions such as image resolution, camera position, occlusion of the face (eg from a disguise, hat, glasses, injury, or hair), lighting,

[169] David White et al, 'Redesigning Photo-ID to Improve Unfamiliar Face Matching Performance' (2014) 20(2) *Journal of Experimental Psychology* 166, 169 ('Redesigning Photo-ID'). See also Clutterbuck and Johnston (n 58) 990. However, this might not apply to super-recognisers: see James D Dunn et al, 'UNSW Face Test: A Screening Tool for Super-Recognizers' (2020) 15(11) *PLoS ONE* e0241747:1–19, 8–9.

[170] Clutterbuck and Johnston (n 58) 990.

[171] This issue was explored in HP Bahrick, PO Bahrick and RP Wittlinger, 'Fifty Years of Memory for Names and Faces: A Cross-Sectional Approach' (1975) 104(1) *Journal of Experimental Psychology* 54. They tested 392 participants' ability to identify classmates from images taken from their high school yearbooks: at 54. Participants were shown 10 randomly chosen images of high school classmates (from cohorts of 200 to 400 students), and asked to identify each person: at 61. Accuracy in this study was 58% (4.5 years after graduation), 51% (7.5 years) and 37% (14.5 years). This is compared to 67% accuracy when tested three months after graduation: at 62.

[172] Ibid. The accuracy rates discussed in that study may be particularly low because participants were asked to correctly *name* a classmate, rather than just provide some identifying information about them: at 61. This would have reduced accuracy uniformly across the groups and so the pattern of diminishing accuracy over time remains informative.

[173] *Tang* (n 9) 683 [2] (Spigelman CJ); *Morgan* (n 22) 42 [56] (Hidden J); *Honeysett* (n 8) 127 [5] (French CJ, Kiefel, Bell, Gageler and Keane JJ). Low-quality images are often admitted in conjunction with other evidence and are thus before the jury. The admissibility of the images was not contested in *Tang* (n 9), *Morgan* (n 22) and *Honeysett* (n 8), only the opinions about the persons and any similarities.

[174] See, eg, Kristin Norell et al, 'The Effect of Image Quality and Forensic Expertise in Facial Image Comparisons' (2015) 60(2) *Journal of Forensic Sciences* 331, 336–9.

and subject-to-camera distance.[175] Accuracy is also affected by changes in appearance caused by factors relating to the face: for example, expression, ageing, weight change or surgery, and head angle.[176]

The potentially detrimental effect of image quality factors on performance is not limited to lay persons. The accuracy of facial examiners and algorithms (more below) is also impaired by degradation in image quality.[177] Interestingly, Kristin Norell et al showed that specialist facial examiners were better able to moderate the *confidence* of their judgments based on the quality of the image evidence than lay persons.[178] When image quality was poor, facial examiners were more likely than students (ie those without training and socialisation) to use the 'unsure' response option in their assessment of whether faces 'matched'.[179] Courts (and decision-makers) should expect expert witness reports to be accompanied by a careful assessment of image quality factors that constrain the type and strength of conclusions available. We should note that while the accuracy of facial recognition algorithms is markedly impaired by reduction in image quality, the current generation of 'deep learning' algorithms appears capable of accommodating some degree of variation in head angle and image quality.[180]

Face matching accuracy is also influenced by the number of images available for comparison.[181] Research shows that the provision of two reference images rather than one results in a slight improvement in accuracy, up to 7%, particularly when the images being compared depict the same person.[182]

[175] Glenn Porter, 'A New Theoretical Framework regarding the Application and Reliability of Photographic Evidence' (2011) 15(1) *International Journal of Evidence and Proof* 26, 57. See also Eilidh Noyes et al, 'The Effect of Face Masks and Sunglasses on Identity and Expression Recognition with Super-Recognizers and Typical Observers' (2021) 8(3) *Royal Society Open Science* 201169:1–18, 6–13.

[176] Hancock, Bruce and Burton (n 152) 333. The low-quality audiovisual systems used to display images and sounds in courtrooms may be an additional source of error.

[177] Norell et al (n 174) 336.

[178] Ibid. On facial examiners, see below Part III(B)(2).

[179] Norell et al (n 174) 339. Purported experts have a tendency to categorically identify (ie positively identify).

[180] Phillips et al, 'Face Recognition Accuracy' (n 153) 6172.

[181] White et al, 'Redesigning Photo-ID' (n 169) 169.

[182] Ibid 170. Having access to multiple reference images is thought to improve accuracy by providing information on how the appearance of a face varies, along with which characteristics are transient properties of the image rather than stable features of the face: see at 168–9.

3 *Demographics of the Face and Decision-Makers*

Another important consideration is the interaction between the demographic profiles of the person pictured in the image(s) *and* the person (or algorithm) making the face identification decision. An example of such an interaction is the phenomenon known as the 'other-race effect'.[183] We are typically better at identifying faces that share our ethnicity than faces of different and less familiar ethnicities.[184] This effect has been replicated in many studies, most of which test the ability of people to remember faces.[185] In a review of this literature, Christian Meissner and John Brigham found that across 39 studies and nearly 5,000 subjects, people were 1.4 times more likely to recognise own-ethnicity compared to other-ethnicity faces.[186]

While the vast majority of studies demonstrating the 'other-race effect' have used tests of memory,[187] a small number have examined other-race effects in simultaneous matching tasks — like those in Figure 5B above. It is important to know whether race influences simultaneous matching because the decisions made by people admitted as some kind of expert (as opposed to eyewitnesses and familiars) do not typically rely on memory. The individuals undertaking comparisons have simultaneous access to two or more images. The limited evidence available suggests that performance is poorer where the ethnicity of the POI is different to the ethnicity (or experience) of the decision-maker.[188]

Ethnicity is not the only demographic factor warranting consideration. Other studies have shown participants to be better at identifying faces from their own age group.[189] Indeed, prominent theoretical accounts of both the

[183] Christian A Meissner and John C Brigham, 'Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review' (2001) 7(1) *Psychology, Public Policy, and Law* 3, 4.

[184] Ibid 3.

[185] Ibid 5.

[186] Ibid 15. The significance of this issue is accentuated by the fact that accused in the leading face identification cases are not Caucasian: consider *Smith* (n 20) 662–3 [38] (Kirby J); *Tang* (n 9) 687–8 [25] (Spigelman CJ); *Alrekabi* (n 39) 296 [30] (Knox DCJ); *Morgan* (n 22) 46 [80] (Hidden J); *Honeysett* (n 8) 129 [16] (French CJ, Kiefel, Bell, Gageler and Keane JJ).

[187] Meissner and Brigham (n 183) 5.

[188] See Ahmed M Megreya, David White and A Mike Burton, 'The Other-Race Effect Does Not Rely on Memory: Evidence from a Matching Task' (2011) 64(8) *Quarterly Journal of Experimental Psychology* 1473, 1479. Of course, features attributed to race will not always be observable, especially if the 'crime' images are of low quality. In *Smith* (n 20) 662–3 [38] (Kirby J), *Morgan* (n 22) 46 [80] (Hidden J) and *Honeysett* (n 8) 127 [9] (French CJ, Kiefel, Bell, Gageler and Keane JJ), the accused were Indigenous Australians. The anatomist involved in *Morgan* (n 22) and *Honeysett* (n 8), Professor Henneberg, had moved to Australia from Poland.

[189] Matthew G Rhodes and Jeffrey S Anastasi, 'The Own-Age Bias in Face Recognition: A Meta-Analytic and Theoretical Review' (2012) 138(1) *Psychological Bulletin* 146, 164.

'other race' and 'own age' effects propose that they have a common perceptual basis.[190] Proponents argue that our relative ability is driven by our perceptual expertise with specific groups.[191] Because perceptual expertise is derived from our experience, we are more adept at recognising and distinguishing the types of faces that we encounter often in our daily lives — and these tend to be people who share our demographic profile.[192]

Interestingly, research suggests that the accuracy of face recognition algorithms is influenced by the demographics of the datasets used to 'train' them.[193] For example, P Jonathon Phillips et al tested the accuracy of algorithms submitted to the National Institute of Standards and Technology ('NIST') Face Recognition Grand Challenge benchmarking exercise.[194] Algorithms developed in East Asian countries performed better on East Asian faces and algorithms developed in Western countries performed better on Caucasian faces — thereby mimicking the other-race effect observed in human participants.[195] This differential accuracy is thought to be because the East Asian algorithms are trained on predominantly East Asian faces, and the Western algorithms are trained on predominantly Caucasian faces.[196] Current state-of-the-art deep learning algorithms also show differential accuracy for different ethnic groups.[197]

[190] Virginia Harrison and Graham J Hole, 'Evidence for a Contact-Based Explanation of the Own-Age Bias in Face Recognition' (2009) 16(2) *Psychonomic Bulletin & Review* 264, 265.

[191] See, eg, Meissner and Brigham (n 183) 9.

[192] Ibid.

[193] See KS Krishnapriya et al, 'Characterizing the Variability in Face Recognition Accuracy Relative to Race' [2019] *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 2278, 2278; P Jonathon Phillips et al, 'An Other-Race Effect for Face Recognition Algorithms' (2011) 8(2) *ACM Transactions on Applied Perception* 14:1–11, 7–8, 10 ('An Other-Race Effect').

[194] Phillips et al, 'An Other-Race Effect' (n 193) 3. In the NIST scheme, algorithm developers voluntarily submit their algorithms for independent testing. The NIST runs the submitted algorithms through tests they have developed where ground truth (ie the correct answer) is known, and the results are published in publicly available reports.

[195] Ibid 5.

[196] Ibid 7–8, 10.

[197] Jacqueline G Cavazos et al, 'Accuracy Comparison across Face Recognition Algorithms: Where Are We on Measuring Race Bias?' (2021) 3(1) *IEEE Transactions on Biometrics, Behavior, and Identity Science* 101, 101, 109.

Importantly, it is currently unknown whether these effects occur in genuinely expert groups, such as facial examiners and super-recognisers[198] (on these groups, more below).[199]

B *Evaluating Performance of the Different 'Groups'*

Now we turn to describe scientific research on the accuracy of face comparison and identification. We want to indicate at the outset that the divisions (or 'groups') we have employed are not always neat; nevertheless, they provide an indication of relative abilities as well as the need to attend to the formal testing *of individuals* rather than relying on proxies such as job title (or other nomenclature), reputation, experience, training or formal qualifications. The review begins with lay persons. We include those we have described as purported experts (eg ad hoc experts, investigators and reviewers) in this discussion for, as we shall see, the available evidence suggests that they do not typically outperform lay persons.[200] We then consider individuals who have consistently demonstrated superior performance in some tasks and conclude our overview with the latest algorithms (employing artificial intelligence) and new types of hybrid systems.

1 *Lay Persons (including Investigators, Reviewers and Other Purported Experts)*

The accuracy of lay persons (or novices) in unfamiliar face matching tasks is surprisingly poor.[201] When lay persons are given unlimited time to complete

[198] In the absence of evidence to the contrary, the perceptual expertise account predicts that experts will also perform better with the types of faces they encounter most frequently. Preliminary work suggests that super-recognisers experience the other-race effect: see Sarah Bate et al, 'The Limits of Super Recognition: An Other-Ethnicity Effect in Individuals with Extraordinary Face Recognition Skills' (2019) 45(3) *Journal of Experimental Psychology* 363, 363. See also David J Robertson et al, 'Super-Recognisers Show an Advantage for Other Race Face Identification' (2019) 34(1) *Applied Cognitive Psychology* 205, 209, 213.

[199] See below Part III(B)(2)–(3).

[200] This would seem to make their evidence irrelevant on the reasoning advanced by the majority in *Smith* (n 20) 654–6 [9]–[12] (Gleeson CJ, Gaudron, Gummow and Hayne JJ).

[201] This finding came as a surprise to the first generation of scientists studying face perception: see Romina Palermo et al, 'Do People Have Insight into Their Face Recognition Abilities?' (2017) 70(2) *Quarterly Journal of Experimental Psychology* 218, 218–19. One of the explanations given is that we make mistakes with unfamiliar faces because we have very little need to identify unfamiliar people in our daily lives, and thus do not develop these skills: see at 219–20. Further, if we fail to recognise that we have encountered an unfamiliar person before, for example, we do not typically receive feedback so we do not learn from the error: at 220. We encourage readers to undertake some of the many face matching tests available online: see, eg, 'UNSW Face Test', *University of New South Wales* (Web Page) <https://facetest.psy.unsw.edu.au>, archived at <https://perma.cc/PSE65W6J>.

face matching tasks that require them to determine whether two photographs depict the same person (as in Figure 5B), error rates range from around 20% with high-quality standardised images to 30–40% in more challenging tests where images are captured in unconstrained environmental conditions — eg from CCTV cameras or photographs taken from social media such as Facebook.[202] For a variety of reasons, including the general lack of feedback in everyday life (in relation to unfamiliar persons), most of us are oblivious to the error-prone nature of unfamiliar face comparison.[203]

In recent years, attention has turned from testing the accuracy of lay persons towards testing professional staff who perform face comparison tasks as part of their daily work.[204] These staff range from 'reviewers' — those who compare images (or images with persons) as a component of their work — to specialist 'facial examiners' — whose primary work involves image interpretation.[205] Reviewers are responsible for confirming the identity of persons, such as those checking documents at national borders, issuing passports, as well as some police officers and security personnel (eg nightclub bouncers). At the other end of the spectrum are specialists engaged primarily, sometimes exclusively, in the interpretation of images for the purpose of identification. These facial examiners have usually spent many years acquiring their skills, often in boutique law

202 See Bruce et al (n 164) 210–11; Burton, White and McNeill (n 160) 286; Josh P Davis and Tim Valentine, 'CCTV on Trial: Matching Video Images with the Defendant in the Dock' (2009) 23(4) *Applied Cognitive Psychology* 482, 488; Zoë Henderson, Vicki Bruce and A Mike Burton, 'Matching the Faces of Robbers Captured on Video' (2001) 15(4) *Applied Cognitive Psychology* 445, 446; Phillips et al, 'Face Recognition Accuracy' (n 153) 6173. Interestingly, similar levels of error (around 30%) led the US President's Council of Advisors on Science and Technology to discourage the use and admission of bite-mark comparison evidence and to suspend further funding for research in that area: *PCAST Report* (n 99) 86 n 239, 87. The Federal Bureau of Investigation also abandoned the use of bullet-lead analysis and modified its use of microscopic hair comparison because of high error rates and persistent over-claiming by forensic practitioners: see 'FBI Laboratory Announces Discontinuation of Bullet Lead Examinations', *Federal Bureau of Investigation* (Press Release, 1 September 2005) <https://www.fbi.gov/news/pressrel/press-releases/fbi-laboratory-announces-discontinuation-of-bullet-lead-examinations>, archived at <https://perma.cc/5T7H-EH44>; 'FBI Testimony on Microscopic Hair Analysis Contained Errors in at Least 90 Percent of Cases in Ongoing Review', *Federal Bureau of Investigation* (Press Release, 20 April 2015) <https://www.fbi.gov/news/pressrel/press-releases/fbi-testimony-on-microscopic-hair-analysis-contained-errors-in-at-least-90-percent-of-cases-in-ongoing-review>, archived at <https://perma.cc/99L5-3JTR>. On the abandonment of bullet-lead analysis, see also Committee on Scientific Assessment of Bullet Lead and Elemental Composition Comparison, National Research Council, *Forensic Analysis: Weighing Bullet Lead Evidence* (Report, 2004) 109–13.

203 Palermo et al (n 201) 220.

204 See, eg, Alice Towler et al, 'Do Professional Facial Image Comparison Training Courses Work?' (2019) 14(2) *PLoS ONE* e0211037:1–17.

205 Ibid 7, 10.

enforcement environments.[206] There is now a relatively large scientific literature comparing the performance of these professional groups — ie reviewers and facial examiners — to lay persons.[207] A key insight from this research is that merely performing the task of face comparison in daily work does not improve accuracy — see Figures 6 and 7 below.[208] That is, experience — merely doing the same thing over and over, such as routinely comparing photographs in passport applications or comparing photographs with applicants at national borders — does *not* improve accuracy.[209]

In a recent meta-analysis of published tests of unfamiliar face matching performance, on 12 out of 18 tests, reviewers — those who perform face comparison as part of their daily work — performed no better than lay persons.[210] The reviewer groups included staff with titles and roles that might reasonably lead people to expect them to possess genuine abilities in face matching.[211] A test of Australian passport officers, for example, found no correspondence between length of employment and accuracy at face matching.[212] Staff who had been employed as reviewers for decades were no more accurate than those employed for just months (see Figure 6) and passport officers were no more accurate than university students.[213] This finding also extends to memory-based face identification tasks.[214] A review of experimental evidence comparing the accuracy of police officers and lay persons at identifying suspects from line-ups revealed no

[206] The precise mechanisms for acquiring this expertise are uncertain and currently the subject of study. We know that short courses and bare experience are insufficient, so it is postulated that expertise develops via extensive on-the-job training: see Alice Towler, David White, Richard I Kemp, 'Evaluating Training Methods for Facial Image Comparison: The Face Shape Strategy Does Not Work' (2014) 43(2–3) *Perception* 214, 214–16; Towler, White and Kemp, 'Evaluating the Feature Comparison Strategy' (n 153) 48, 56–7; White et al, 'Passport Officers' Errors in Face Matching' (n 153) 3. Facial examiners are discussed in detail below in this Part.

[207] See, eg, Alice Towler, Richard I Kemp and David White, 'Can Face Identification Ability Be Trained: Evidence for Two Routes to Expertise' in Markus Bindemann (ed), *Forensic Face Matching: Research and Practice* (Oxford University Press, 2021) 89, 100–5.

[208] It may be a necessary component but alone is insufficient.

[209] See White et al, 'Passport Officers' Errors in Face Matching' (n 153) 1. This would seem to suggest that comparing faces as part of a prior investigation will have no significance in terms of performance.

[210] White, Towler and Kemp (n 153) 69.

[211] Ibid.

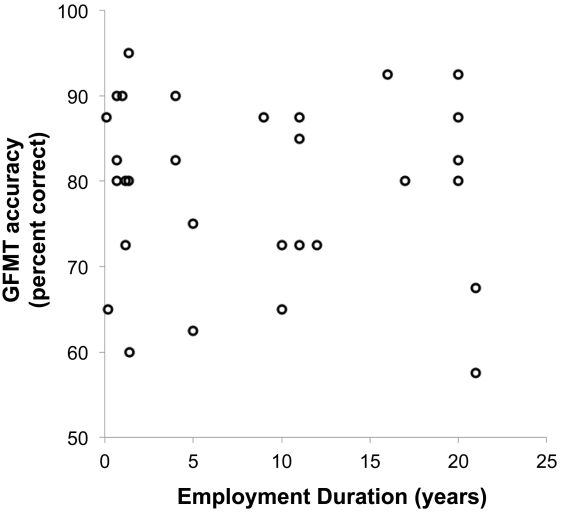
[212] White et al, 'Passport Officers' Errors in Face Matching' (n 153) 3.

[213] White, Towler and Kemp (n 153) 69–70. This finding seems to be consistent with studies of fingerprint examiners, where length of time in the job (ie experience), after an initial training period, does not correspond with improved accuracy: see Rachel A Searston and Jason M Tangen, 'The Emergence of Perceptual Expertise with Fingerprints over Time' (2017) 6(4) *Journal of Applied Research in Memory and Cognition* 442, 448–9.

[214] See White et al, 'Passport Officers' Errors in Face Matching' (n 153) 3.

difference between these groups in five of seven studies.[215] Together, this evidence shows that professional role and experience in face identification tasks are no guarantee of enhanced ability — ie expertise.

Figure 6: Passport officers' accuracy on a standard test of face matching ability (Glasgow Face Matching Test ('GFMT')), plotted against years of experience. Each dot represents a passport officer. Accuracy does not improve with experience.[216]



The fact that experienced reviewers perform no better than persons without experience also raises questions about the value of training.[217] Alice Towler et al tested reviewers before and after they received professional training in face comparison.[218] The courses reviewed promoted detailed analysis of facial features, instruction on facial features, and components of features (eg the tragus, helix, antihelix, and lobe of the ear). The content tended to be consistent with international recommendations — eg the best-practice guidelines set out by the

[215] See Annelies Vredeveldt and Peter J van Koppen, 'The Thin Blue Line-Up: Comparing Eyewitness Performance by Police and Civilians' (2016) 5(3) *Journal of Applied Research in Memory and Cognition* 252, 254. However, in one study police made *more* false positive identifications (ie errors) than lay people.

[216] Figure reproduced from White et al, 'Passport Officers' Errors in Face Matching' (n 153) 2.

[217] It also raises questions about what distinguishes the training and experience of high-performing facial examiners from reviewers.

[218] Towler et al (n 204) 4; White, Towler and Kemp (n 153) 66–8.

Facial Identification Scientific Working Group ('FISWG').[219] Nevertheless, Towler et al found little evidence that formal training improves performance.[220] This suggests that training (like experience) does not simply manifest in a heightened level of ability — ie expertise.[221]

By now the reader should be increasingly sensitive to the value of formally evaluating performance — in order to generate knowledge of abilities and accuracy — rather than assuming the existence of expertise on the basis of 'training, study or experience' or imputing 'knowledge'. In investigations and criminal proceedings we should expect (and be presented with) *empirical evidence* of the actual performance of persons said to be expert and algorithms said to be accurate.[222]

Fortunately, individuals and algorithms with genuine abilities are (now) known to exist.

2 *Facial Examiners — (Some) Genuine Expertise in Image Comparison*

In their recent review of tests that compared the accuracy of professional groups to lay persons, David White, Alice Towler and Richard Kemp found that facial examiners — specialists who are primarily engaged in image interpretation — consistently outperformed lay persons and reviewers (such as passport examiners).[223] Facial examiners are typically employed by government departments (eg the Department of Foreign Affairs and Trade) or policing agencies (eg the Federal Bureau of Investigation) and spend most, sometimes all, of their time interpreting images.[224] They are routinely involved in making decisions about the identity of POIs and may spend hours, days or even weeks on particular

[219] Facial Identification Scientific Working Group, *Guidelines and Recommendations for Facial Comparison Training to Competency* (Guidelines Version 1.1, 18 November 2010); Towler et al (n 204) 2–4. FISWG is primarily an industry body composed of those engaged in facial comparison work (dominated by US agencies) and has mixed engagement with scientific research. Almost none of its documents is scientifically based: see 'FISWG Documents', *Facial Identification Scientific Working Group* (Web Page) <http://fiswg.org/documents.html>, archived at <https://perma.cc/JFW4-A2B5>.

[220] Towler et al (n 204) 9–10.

[221] This raises questions about 'training' and 'experience' under s 79(1) of the *Uniform Evidence Law* (n 6) because, in practice, either might be used to support the admission of opinions (without attending to specialised knowledge), particularly after *Dasreef* (n 40) 604 [37] (French CJ, Gummow, Hayne, Crennan, Kiefel and Bell JJ): see Edmond and Martire, 'Knowing Experts' (n 40) 88–90.

[222] Lawyers, judges and jurors should not be left to make guesses about performance, especially when these things can be, and have been, tested.

[223] White, Towler and Kemp (n 153) 68, 75.

[224] The Australian representatives tend to produce 'investigative' information, although some have prepared reports in relation to immigration and refugee matters and some have presumably been used in cases where pleas were accepted: see ibid 74.

comparisons. Examiners often provide written reports that explain the bases for their identification decisions and in some jurisdictions (though not currently Australia) submit these reports to courts and, if required, provide expert testimony. In the seven tests conducted so far, facial examiners consistently outperformed lay persons.[225] These differences were typically large, ranging from 10% to 20% higher accuracy.[226]

Facial examiners *as a group* have demonstrated the kind of enhanced ability conventionally associated with expertise.[227] However, the number of studies is modest.[228] The small number of tests conducted so far use images of 'compliant' subjects looking directly at a camera, in relatively good lighting and with good resolution (as in Figure 5B).[229] Scientists are yet to conduct systematic tests of the effects of image quality factors on the performance of facial examiners. However, preliminary evidence suggests that image quality will be as detrimental to their performance as it has been to the accuracy of lay persons.[230]

Importantly, the tests conducted so far reveal large inter-examiner variation.[231] That is, there are substantial differences between the performances of different facial examiners. While facial examiners perform very well as a group (in comparison to other groups such as reviewers and lay persons), some individual facial examiners perform poorly.[232] The results from the initial studies should not be extrapolated to all of those who are, or claim to be, facial examiners. In the largest test conducted so far, Phillips et al found that the performance of individual facial examiners spanned the entire range of the measurement scale — from near chance level (ie 50% — expected from guessing) to perfect accuracy (ie 100%).[233] Revealingly, seven of 57 facial examiners made errors in more than 30% of their comparisons.[234] All of the examiners tested were members of internationally respected scientific working groups and they

225 Ibid 75. These are important comparisons because they demonstrate that the examiners have actual expertise, rather than just claimed or attributed (ie purported) expertise.

226 Ibid.

227 See ibid.

228 Though presumably sufficient for admission: see Kemp, Edmond and White (n 3) 17. Even jurisdictions with reliability standards do not require much.

229 See White, Towler and Kemp (n 153) 62, 76.

230 Norell et al (n 174) 336.

231 See, eg, Phillips et al, 'Face Recognition Accuracy' (n 153) 6172.

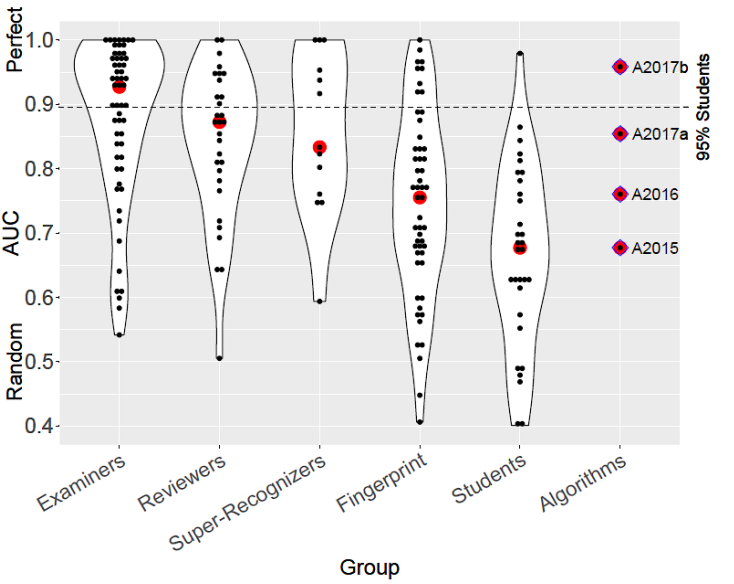
232 Ibid 6173.

233 Ibid.

234 See Figure 7. These particular examiners should not be relied upon when there are much more capable examiners available, as the figure illustrates.

were allotted three months to complete 20 same–different face matching decisions — the type of task depicted in Figure 5B.[235]

Figure 7: Face identification/matching accuracy of forensic examiners, reviewers, super-recognisers and algorithms. Small black dots denote one person's/algorithm's performance, and large dots denote group medians. The white 'violins' denote the distribution and density of scores. The black dotted line indicates the performance of 95% of students, such that scoring above this line indicates a person performs better than 95% of students. The most recent algorithms — the years of development are included in their titles — performed as well as the very best humans.[236]



The studies reviewed by White, Towler and Kemp provide evidence that some professional groups attain levels of accuracy that would seem to satisfy admissibility criteria concerned with genuine expertise or knowledge. However, the preliminary nature of these studies means that caution should be exercised when applying the group findings to any specific individual who might be presented as an expert witness.[237] It seems important to have a clear idea of

[235] See Phillips et al, 'Face Recognition Accuracy' (n 153) 6172, 6175–6.

[236] Figure taken from ibid 6173.

[237] That is, scientific research confirms that many facial examiners possess genuine expertise in unfamiliar face matching. In consequence, their opinions would seem to be based on this 'specialised knowledge': *Uniform Evidence Law* (n 6) s 79(1).

individual performance — based on standardised testing in controlled conditions — rather than to extrapolate from the average performance of groups, job title, training or years of experience. Demonstrable personal ability would seem to be an essential prerequisite for admission, reliance and comprehension.[238]

3 *Super-Recognisers — Ability without Knowledge, Training, Study or Experience*

Relatively recently, while studying prosopagnosia (ie face blindness), scientists identified a group of individuals whose abilities might provide an alternative (or supplement) to facial examiners.[239] These are individuals who seem to have a natural aptitude for face matching. In the scientific literature and the media they have been labelled 'super-recognisers'.[240] They represent the very top end of a naturally varying continuum of face identification ability across the general population. A large body of evidence confirms there is substantial variation between people's abilities in face comparison tasks.[241] This variation is 'relatively stable across repeated testing'.[242] Recruiting and selecting forensic practitioners on the basis of their innate face matching ability is therefore a promising way to improve the accuracy of face identification decisions made in professional settings.[243]

With the discovery of variable performance across the general population, in conjunction with the outstanding performance of a handful of police officers reviewing CCTV images following the London riots in 2011, London's

238 Otherwise, courts might unwittingly be admitting the opinions of examiners whose assessments are no better than chance or the jury.

239 Richard Russell, Brad Duchaine and Ken Nakayama, 'Super-Recognizers: People with Extraordinary Face Recognition Ability' (2009) 16(2) *Psychonomic Bulletin and Review* 252, 254.

240 See, eg, ibid; Phillips et al, 'Face Recognition Accuracy' (n 153) 6171.

241 See ibid 6173.

242 Tarryn Balsdon et al, 'Improving Face Identification with Specialist Teams' (2018) 3 *Cognitive Research: Principles and Implications* 25:1–13, 1. See also Jeremy B Wilmer et al, 'Human Face Recognition Ability Is Specific and Highly Heritable' (2010) 107(11) *Proceedings of the National Academy of Sciences* 5238, 5239. For a review, see Jeremy B Wilmer, 'Individual Differences in Face Recognition: A Decade of Discovery' (2017) 26(3) *Current Directions in Psychological Science* 225. Cf Meike Ramon, Anna K Bobak and David White, 'Super-Recognizers: From the Lab to the World and Back Again' (2019) 110(3) *British Journal of Psychology* 461, 474–5.

243 White et al, 'Passport Officers' Errors in Face Matching' (n 153) 5–6; Anna Katarzyna Bobak, Andrew James Dowsett and Sarah Bate, 'Solving the Border Control Problem: Evidence of Enhanced Face Matching in Individuals with Extraordinary Face Recognition Skills' (2016) 11(2) *PLoS ONE* e0148148:1–13, 10. The idea, raised by Heydon J in *Aytugrul* (n 149) 203 [75], that a member of the jury might be reasonably skilled in a particular area, in this case, being good with faces (in *Aytugrul* (n 148), it was with mathematics) would not seem to be a solution, for individuals with superior abilities may not know about their abilities relative to others, and may or may not exert a positive influence on jury deliberations.

Metropolitan Police Service ('Met') established a 'super-recogniser' team in 2013.[244] This team seems to have been an amalgam of those who identified persons with whom they were familiar from localities in which they worked (eg custody officers and local police who identified recidivists) and/or those who performed well on standardised face matching and face memory tests (eg the GFMT and the Cambridge Face Memory Test).[245] Notwithstanding the large number of studies showing variation in face matching tasks across the general population, few studies have tested professional groups of super-recognisers.[246] The only evidence comes from two studies comparing the accuracy of super-recognisers employed by the Met with untrained novices.[247]

Josh Davis et al tested 36 Met super-recognisers on a standard face matching test.[248] Super-recognisers outperformed university students on this test by six percentage points.[249] In a similar study, David Robertson et al tested four super-recognisers from the Met on the same test and found they outperformed police trainees by 17.8%, and outperformed a control group of university undergraduates by 22.7% and 27% on two other challenging face identification tasks.[250] It is not entirely clear why this group of four performed so much better than the group of 36, but we suspect these four super-recognisers were a subset of the staff tested by Davis et al, and were selected to form an elite team on

[244] See Alex Moshakis, 'Super Recognisers: The People Who Never Forget a Face', *The Guardian* (online, 11 November 2018) <https://www.theguardian.com/uk-news/2018/nov/11/super-recognisers-police-the-people-who-never-forget-a-face>, archived at <https://perma.cc/D33M-X53F>.

[245] See David J Robertson et al, 'Face Recognition by Metropolitan Police Super-Recognisers' (2016) 11(2) *PLoS ONE* e0150036:1–8, 2. See also the tests and links available at *UNSW Sydney Forensic Psychology Lab* (Web Page) <http://forensic.psy.unsw.edu.au>, archived at <https://perma.cc/5E3N-VG9J>.

[246] Super-recognisers are also discussed below in Part III(B)(4). The super-recognisers included in the tests recorded in Figure 7 were not all practising in a professional capacity and may not represent the very top levels of facial comparison and memory ability: see Phillips et al, 'Face Recognition Accuracy' (n 153) 6172.

[247] Josh P Davis et al, 'Investigating Predictors of Superior Face Recognition Ability in Police Super-Recognisers' (2016) 30(6) *Applied Cognitive Psychology* 827; Robertson et al, 'Face Recognition by Metropolitan Police Super-Recognisers' (n 245).

[248] Davis et al (n 247) 829. Participants completed the GFMT discussed in Burton, White and McNeill (n 160) 287–8. Standard face matching tests are variations on Figure 5B. Participants are asked to decide whether two (or more) images, usually shown simultaneously on a computer screen, are of the same person or different people: at 286.

[249] Davis et al (n 247) 835.

[250] Robertson et al, 'Face Recognition by Metropolitan Police Super-Recognisers' (n 245) 4–6.

the basis of on-the-job performance in conjunction with high scores on standardised tests.[251]

Exactly how super-recognisers achieve such high levels of accuracy is not yet clear — super-recognisers were only 'discovered' in 2009.[252] Since that time, researchers have tended to focus on verifying their superior ability rather than studying *how* they achieve it.[253] Consequently, we do not yet understand how such accurate identification decisions are produced.

Of interest, preliminary evidence suggests that super-recognisers lack the conservatism of facial examiners.[254] That is, they are more likely to make errors with high levels of confidence.[255] This aspect of performance may follow from being asked questions about whether faces match in ways that are quite abstract. Facial examiners, in contrast, are more likely to be sensitive to forensic uses and institutional values — including the risks of overstatement and error.[256]

4 *Algorithms*

The use of deep learning has improved face recognition algorithms dramatically in recent years.[257] The accuracy of these algorithms has reached the stage where they are superior to the average lay person in making unfamiliar face matching decisions.[258] In a recent comparison between algorithms, facial examiners, reviewers, super-recognisers and lay people, Phillips et al found that algorithms perform just as accurately as the best performing humans (ie facial examiners and super-recognisers).[259] In this particular study, 'participants' were required to determine whether sets of two photographs depicted the same or different persons.[260] The most reliable algorithm (A2017b) was 96% accurate.[261] On average, facial examiners were 93% accurate and super-recognisers were

[251] See Davis et al (n 247) 829. It may be that some officers performed well because they were reasonably familiar with many of the 'usual suspects' they had arrested or supervised as custody officers: at 837.

[252] Russell, Duchaine and Nakayama (n 239) 254.

[253] See, eg, Phillips et al, 'Face Recognition Accuracy' (n 153) 6171–2.

[254] See ibid 6174.

[255] Ibid.

[256] White, Towler and Kemp (n 153) 78, citing Norell et al (n 174) 337–8. Whether this is desirable in experts is a separate question. Further, super-recognisers may be able to learn to become more conservative in their opinions.

[257] See Phillips et al, 'Face Recognition Accuracy' (n 153) 6172–3.

[258] See ibid.

[259] Ibid.

[260] Ibid 6172.

[261] Ibid.

83% accurate.[262] All three groups — algorithms, facial examiners and super-recognisers — outperformed lay people, comprising fingerprint examiners (76%) and students (68%).[263] However, algorithms and super-recognisers were much faster than facial examiners.[264]

Figure 7 illustrates fundamental information about the performance of different groups in a way that helps us understand the value of their interpretations. Small black dots denote the accuracy of a single person/algorithm, and the large dots denote average performance of each group. Such studies provide the kind of knowledge that should inform decisions about the admission and use of images as well as the probative value of opinions about images for purposes of identification. It provides a clear indication of the types of face matching opinions that tend to be (most) probative, while also providing an estimate of the degree of variance in accuracy that can be expected within each of these types.[265] Courts, as we explain below in Part IV, should be primarily focused on those whose superior abilities with face comparisons have been empirically demonstrated. These are largely found among facial examiners and super-recognisers, along with the increasingly sophisticated algorithms.

5 *Hybrid Systems (and Meta-Analysts)*

Up until this point we have discussed the performance of various humans and face recognition algorithms in isolation. However, in most 'real-world' settings face identification decisions are made by hybrid systems, in which humans and algorithms (usually software) work in combination.[266] For example, face recognition software is most often used to search large national image databases (eg of criminal mugshots or drivers' licences) to find faces that look most similar to a photograph of a POI.[267] A human analyst must then interrogate the search

[262] Ibid 6172–3. Super-recognisers performed below the level of facial examiners and some algorithms, but this is likely a consequence of fairly lax criteria for inclusion in this super-recogniser group compared to other studies.

[263] Ibid 6172. The inclusion of fingerprint examiners confirms that expertise is not simply transferable. Being a fingerprint examiner does not translate into superior performance at face comparison.

[264] Ibid.

[265] Reviewers tested in this study perform more specialised work than other groups of reviewers in the literature, which may explain their relatively high performance on this task: see Phillips et al, 'Face Recognition Accuracy' (n 153) 6172.

[266] Alice Towler, Richard I Kemp and David White, 'Unfamiliar Face Matching Systems in Applied Settings' in Markus Bindemann and Ahmed M Megreya (eds), *Face Processing: Systems, Disorders and Cultural Differences* (Nova Science Publishers, 2017) 21, 23–4 ('Unfamiliar Face Matching Systems').

[267] Ibid 23; Phillips et al, 'Face Recognition Accuracy' (n 153) 6172.

results to decide if the POI is present.[268] Studies confirm that combining humans and algorithms in this way drastically *increases* errors.[269] White et al found that passport reviewers — those who perform this very task as a component of their daily work — made errors more than 50% of the time when adjudicating the output of face recognition software.[270] This high error rate is an unintended consequence of face recognition software finding the most similar faces in a database containing millions of images.[271] The faces presented to the human analyst are often extremely similar.

Empirical research reinforces the need to carefully design hybrid systems based on known abilities and procedures that reduce the risk of error. One particularly effective method of improving accuracy is to employ the 'wisdom of crowds' or fusion.[272] In these systems humans and algorithms work in parallel rather than serially, as described above. Humans and/or algorithms each make *independent* decisions which are then aggregated, usually by averaging the scores, to produce a single 'group' decision.[273] The group decision is often more accurate than any of the individual scores that produced it.[274] Phillips et al recently demonstrated that combining the decision of a single facial examiner and the best available algorithm often resulted in perfect performance.[275] Together, these findings illustrate the critical importance of designing evidence-based face identification systems — get it wrong and the results are catastrophic, but get it right and the results are highly accurate and hugely beneficial to fact-finding.

One of the benefits of these hybrid systems is that risks from many biases can be managed and/or dispersed.[276] Additionally, humans and algorithms can

[268] Towler, Kemp and White, 'Unfamiliar Face Matching Systems' (n 266) 23. This is similar to the way algorithms (eg the Automated Fingerprint Identification System) are applied to assist with fingerprint comparison in most policing agencies: see Gary Edmond, 'Fingerprint Evidence in New Zealand's Courts: The Oversight of Overstatement' (2020) 29(1) *New Zealand Universities Law Review* 1, 18 ('Fingerprint Evidence in New Zealand's Courts').

[269] David White et al, 'Error Rates in Users of Automatic Face Recognition Software' (2015) 10(1) *PLoS ONE* e0139827:1–14, 10. Cf Phillips et al, 'Face Recognition Accuracy' (n 153) 6172.

[270] White et al, 'Error Rates in Users of Automatic Face Recognition Software' (n 269) 10.

[271] Ibid 11.

[272] James Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (Little, Brown, 2004) xiv; David White et al, 'Crowd Effects in Unfamiliar Face Matching' (2013) 27(6) *Applied Cognitive Psychology* 769, 769.

[273] Phillips et al, 'Face Recognition Accuracy' (n 153) 6171; Towler, Kemp and White, 'Unfamiliar Face Matching Systems' (n 266) 29–30.

[274] Phillips et al, 'Face Recognition Accuracy' (n 153) 6174.

[275] Ibid.

[276] See ibid; Towler, Kemp and White, 'Unfamiliar Face Matching Systems' (n 266) 30.

be tested on image pairs where the correct answer (ie ground truth) is known.[277] This testing can be imposed so that the persons involved in comparisons do not know that any particular task is a test.[278] It can appear to be part of their routine flow of work. Such testing provides very useful information about the performance of the system. With both algorithms and hybrid systems we can obtain and provide useful information about their accuracy in conditions similar to those in the investigation or prosecution.[279] This information can be provided to those considering admission and evaluating the outputs.

Algorithms and hybrid systems will introduce new questions: whether the output should be treated as opinion evidence and/or who should be called to explain or account for the output.[280] These are not necessarily novel questions, but in responding, courts should be aware that even if humans are involved in decision-making — whether as a facial examiner using an algorithm as a preliminary search tool or a hybrid system that combines the performances of multiple algorithms and/or humans — it may be desirable to receive a report from, and have questions about the system and limitations addressed by, those who understand how it works. These individuals may not possess special abilities in face comparison and are unlikely to be involved in decisions about whether faces are believed to match (or not). Rather, they are a kind of *meta-expert*. They are capable of explaining the system, the results and the way the results are expressed (based on statistics), in a way that the individual decision-makers — whether facial examiners, super-recognisers, or algorithms — typically cannot.[281]

IV Discussion

Following this review of scientific research, *Tang*, *Honeysett* and *IMM* seem, to put it mildly, inadequate. There appear to be deep, and perhaps structural, impediments to common law legal institutions productively engaging with mainstream scientific knowledge.[282] These impediments seem pronounced in

[277] See, eg, Phillips et al, 'Face Recognition Accuracy' (n 153) 6172.

[278] For example, in Itiel E Dror, David Charlton and Ailsa E Péron, 'Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications' (2006) 156(1) *Forensic Science International* 74, 76, the experts did not know that they were being tested.

[279] These will presumably vary with the quality and quantity of material.

[280] Roth (n 5) 1976–7.

[281] For examples of these kinds of meta-experts see *Tuite* (n 80) 229–34 [115]–[120] (Maxwell ACJ, Redlich and Weinberg JJA) and the discussion in Simon A Cole, 'A Cautionary Tale about Cautionary Tales about Intervention' (2009) 16(1) *Organization* 121, 123–6. See also Brewer (n 79) 1627.

[282] See Edmond, Hamer and Cunliffe (n 149) 410–11.

criminal proceedings.[283] Our courts have placed reliance on legal tradition, judicial experience, and case-based reasoning. They focus attention on legal metaphysics (around fact and opinion) and vest confidence in rules of evidence and procedures, trial safeguards and jury decision-making.[284] Yet, jurisprudence and practice seem insensitive to pertinent scientific research on the actual performance of those allowed to proffer opinions, as well as the abilities of jurors.[285] *None of the research discussed in Part III has been cited in a published legal decision.*[286] Scientific knowledge does not appear to have informed the interpretation of rules or the operation of criminal procedures and trial safeguards said to be part of our rational tradition of proof.[287]

At this juncture it is our intention to draw attention to 10 issues that arise from the foregoing reviews. Some of these points are straightforward, others require wholesale reconsideration of not only our admissibility jurisprudence, but some of the assumptions and procedures associated with our courts' rather cavalier approach to admitting and using images to identify POIs in criminal proceedings.

First, we need to accept that image interpretation is difficult and surprisingly error-prone.[288] When it comes to strangers, we humans regularly make mistakes even in favourable conditions with high-quality images.[289] Furthermore, unfamiliar face matching is not something that we often engage in, and we rarely receive feedback on the accuracy of our unfamiliar face matching

283 They seem to be accentuated by widespread trust in the state and its agencies and any experts they call. Problems are compounded by the lack of resources available to defendants.

284 See, eg, *Haidari v The Queen* (2015) 251 A Crim R 422, 439 [76]–[77], 441 [87]–[89] (Johnson J).

285 Interestingly, legal institutions receive remarkably little independent feedback on their performance. That is, where the correct answer is known. Convictions and appeals do not necessarily reflect ground truth/objective reality and for a variety of reasons, including the restricted involvement of appellate courts, wrongful convictions are rarely exposed *and* recognised.

286 See the decisions discussed above in Part II.

287 Consider the failure to engage with relevant scientific research and norms discussed in Gary Edmond, 'Forensic Science and the Myth of Adversarial Testing' (2020) 32(2) *Current Issues in Criminal Justice* 146, 147–8.

288 See, eg, Glenn Porter, 'Visual Culture in Forensic Science' (2007) 39(2) *Australian Journal of Forensic Sciences* 81, 82–4; Janina Struk, *Photographing the Holocaust: Interpretations of the Evidence* (IB Tauris, 2004) 4, 15. See Errol Morris, *Believing Is Seeing: Observations on the Mysteries of Photography* (Penguin Press, 2011) 112–14; John Berger, 'Understanding a Photograph' in Alan Trachtenberg (ed), *Classic Essays on Photography* (Leete's Island Books, 1980) 291, 293–4; John Tagg, *The Burden of Representation: Essays on Photographies and Histories* (Macmillan Education, 1988) 2–5. For an example of institutional sensitivity with using photographic evidence, see *Goode v Angland* (2017) 96 NSWLR 503, 522 [89]–[93] (Beazley P).

289 See, eg, White et al, 'Passport Officers' Errors in Face Matching' (n 153) 3–4.

abilities.[290] As a result, we dramatically underestimate the difficulty of the task and the number of mistakes we make.[291] We should not assume, therefore, that comparison and identification are tasks that the jury can somehow manage or that trial safeguards or the other evidence will operate as correctives. Decades of scientific research on human cognition confirm that contextual information (whether the opinion of a purported expert or other evidence, such as a motive), and even legal procedures themselves (eg production by the prosecutor and the fact of admission which both imply evidentiary value), are likely to bias the perceptions of those earnestly trying to compare and identify POIs in images.[292] There is, in addition, no evidence that non-independent groups (eg jurors deliberating in common) avoid these vulnerabilities.[293]

Secondly, we should treat the evidence of *anyone* who is identifying a person on the basis of images as opinion evidence.[294] (This includes eyewitnesses, although this article is not substantially engaged with this type of evidence.)[295] It does not matter if this evidence is from a spouse of 50 years or a complete stranger.[296] We should abandon classifying some interpretations as 'fact' or 'direct' or 'recognition' evidence to avoid the implications of all identifications being opinion (and subject to s 76 of the *Uniform Evidence Law*). Identifying persons in images, however quick or intuitive, is always interpretive.[297] It is always

[290] Ibid 5.

[291] Ibid; Palermo et al (n 201) 220.

[292] See Gary Edmond et al, 'Contextual Bias and Cross-Contamination in the Forensic Sciences: The Corrosive Implications for Investigations, Plea Bargains, Trials and Appeals' (2014) 14(1) *Law, Probability and Risk* 1, 16–20 ('Contextual Bias and Cross-Contamination in the Forensic Sciences').

[293] See ibid 17–18, 23–4.

[294] We recognise that all perception is actually interpretive and therefore sits awkwardly within the legal metaphysics around the fact–opinion dichotomy. Here, we are deliberately limiting any extension to identification (and opinions said to be expert). We accept that for most forms of perception and interpretation we do not have better means than the impressions of those who directly perceived the events. Where, however, there is a recording and we have access to genuine expertise that enables us to interpret more accurately than a judge or jury, we should draw upon such opinions and outputs. Consider Frederick Schauer and Barbara A Spellman, 'Is Expert Evidence Really Different?' (2013) 89(1) *Notre Dame Law Review* 1, 8–10.

[295] Of course, there should be exceptions for the opinions of eyewitnesses as well as those who are demonstrably better than jurors. Eyewitnesses and non-investigative familiars should be able to testify.

[296] Cf *Leung* (n 54) 414 [43] (Simpson J); *Nguyen* (n 32) 411–13 [20]–[27] (Basten JA).

[297] See Edmond et al, 'Law's Looking Glass' (n 1) 337. We never really know how long it takes because this is rarely recorded, though witnesses frequently insist that their identifications were instantaneous: see, eg, *Murdoch* (n 56) 340 [70] (Angel ACJ, Riley J and Olsson AJ), citing *R v Williams* (1983) 2 VR 579, 582 (Gobbo J). Sometimes witnesses are told, or it is obvious from the circumstances, who they are there to 'identify': see, eg, *Morgan* (n 22) 44 [72] (Hidden J).

opinion, though we now know that some of these opinions are markedly more reliable than others.[298]

Based on available scientific research and our shared experience, there should be an exception to s 76 for the opinions of non-investigative familiars (where that familiarity is derived from pre-investigative contexts and quotidian interactions over time and across a variety of settings). Such familiars are more accurate than investigators and more accurate than jurors, juries and judges.[299] If the opinions of non-investigative familiars are not clearly accommodated within the existing exceptions to s 76, then a new section should be created for them. As drafted, ss 78–9 are poorly suited to this class of indirect witness.[300] Notwithstanding recent attempts to expand its application to displaced viewers and listeners, this is not an answer to the gap in relation to non-investigative familiars. Section 78 should only confer an exception on those who directly perceived a 'matter or event' — such as ear- or eyewitnesses.

There are compelling reasons for receiving the opinions of those who directly perceive a matter or event (eg eyewitnesses) as well as non-investigative familiars. The opinions of eyewitnesses will often be 'necessary'. The opinions of genuine familiars are probative because of their enhanced ability, as a class of witness, relative to jurors.[301] This is not an ability that jurors are likely to acquire during the course of proceedings.[302] Persons who do not satisfy these conditions are not direct witnesses and should only be able to proffer opinions if they possess demonstrable abilities — they can testify via s 79(1).

This brings us to the *third* point. It concerns the reliability of opinions said to be based on 'specialised knowledge' admitted in the context of a trial. Any individual who testifies about the identity of a POI as an expert should —

[298] See, eg, Phillips et al, 'Face Recognition Accuracy' (n 153) 6173.

[299] Edmond and San Roque (n 51) 30. Non-investigative familiars are generally accurate, though hardly infallible. Consider the issues which arise from cases such as *Polyukhovich v Commonwealth* (1991) 172 CLR 501. Discussion of the evidentiary issues in that case occurred in subsequent proceedings before the Supreme Court of South Australia: see *R v Polyukhovich* (Supreme Court of South Australia, Cox J, 22 December 1992).

[300] That is, those making their identifications on the basis of their memory of a specific and familiar individual.

[301] Edmond and San Roque (n 51) 30–2. We can rely on studies which confirm the heightened abilities of this class of witness (ie familiars) and, in the absence of specific reasons (eg because of some issue with a particular witness's sensory perception), do not need to go to the expense and inconvenience of ascertaining individual abilities. Where an individual holds themselves out, or is held out, as some kind of expert (or to have an advantage, such as ad hoc experts), we should be provided with evidence of individual proficiency to confirm that their opinion is actually relevant, for there is limited support for the assumption that the conditions in which ad hoc expertise is obtained confer an advantage.

[302] This would satisfy the requirements from *Smith* (n 20) 663–4 [41] (Kirby J).

regardless of whether they describe similar features, speak probabilistically, or categorically identify — be demonstrably better than ordinary persons (eg jurors). Procedures and proficiency should have been formally evaluated so that the individual's abilities are *known*, disclosed and able to be considered.[303] For feature comparison forensics, we should not rely on legal tradition, past practice, job titles, self-serving claims, popular beliefs, plausibility, or weak admissibility traditions as proxies for expertise.[304] We should not (have to) assume that persons presented as experts possess genuine abilities.[305] Relevance (and probative value) should not be left to what individual juries might 'accept', nor should it be taken at some speculative 'highest' value.[306]

Where the images are of reasonable quality, facial examiners, super-recognisers and algorithms consistently transcend the performances of ordinary persons.[307] Their opinions and outputs should be obtained and admitted.[308] Where images are of low quality, as in *Tang*, or POIs are well disguised, as in *Honeysett*, we do not currently possess reliable means of identification.[309] That is, there are no relevant experts. Any opinions and outputs are speculative — they are not based on knowledge. This is a lacuna that jurors cannot be expected to fill. As noted above in our first point, they are ill-suited to undertaking such

303 This is consistent with the definition of 'knowledge' in *Honeysett* (n 8) — 'as from study or investigation': at 131–2 [23] (French CJ, Kiefel, Bell, Gageler and Keane JJ).

304 Identification by bite-mark comparison is a good example of a facially plausible type of comparison by highly qualified individuals that in reality offers very limited insight. The President's Council of Advisors on Science and Technology ('PCAST') concluded that 'available scientific evidence strongly suggests that examiners cannot consistently agree on whether an injury is a human [bite mark] and cannot identify the source of [the bite mark] with reasonable accuracy': *PCAST Report* (n 99) 87. The historical use of bite-mark evidence allows us to reflect on the fact that individuals with formal qualifications and experience as dentists are not experts at comparing bite patterns/marks in order to identify POIs: see at 83–7.

305 Juror beliefs should not be capable of transcending or displacing uncontradicted scientific research. Jury confidence in a purported expert, such as a passport examiner or investigative familiar, for example, does not make their opinions relevant or accurate. See also the seventh point raised below in this Part.

306 From time to time, judges suggest that the abilities of purported experts cannot be realistically tested. In *Li* (n 31) 287–90 [44]–[63] (Ipp JA, Whealy J agreeing at 298 [137], Howie J agreeing at 298 [138]), the Court indicated that testing was an unreasonable expectation, yet the *NRC Report* (n 148) 53 and *PCAST Report* (n 99) 6 explain that such testing is not only feasible but necessary.

307 Familiars, and perhaps super-recognisers and facial examiners, appear to have an advantage over algorithms with lower-quality images: see above Part III(A)–(B).

308 This should be subject to the provision of performance information, as described in the fourth point raised in this Part.

309 See Norell et al (n 174) 336; Davis and Valentine (n 202) 494.

demanding face comparisons, especially in suggestive conditions.[310] We should not expect jurors to do what genuine experts cannot.[311]

Fourthly, the opinions of an individual admitted as an expert should always be accompanied by the best estimate of accuracy based on the results of formal empirical evaluation — whether validation studies or rigorous proficiency testing.[312] That is, we should *know* about the particular individual's performance.[313] Expert opinions should be presented with an *indication* of the risk of error based on the individual's performance in (roughly) similar conditions.[314] There is limited value in allowing the bare description of features or categorical identification if decision-makers are not presented with some idea of the witness's ability at the specific task.[315]

Fifthly, any algorithm relied upon in criminal proceedings should have been rigorously evaluated in (roughly) similar conditions. Issues of race, age, the quality and quantity of the images, the role of individuals in the operation of the system, indicative error rates, and so on and so forth, should all be

[310] In general, this evidence should not be adduced and admitted and jurors should not be expected or allowed to undertake their own error-prone comparisons. In many cases, if the defence were to stipulate that the defendant appears similar to the person in the images, then in most cases nothing can be achieved *in terms of identification* by showing low-quality images to the jury.

[311] Here, the existence of other evidence may unconsciously bias jurors when they are invited to compare unfamiliar persons in conditions where genuine experts and the best algorithms are conspicuously error-prone: see the first point raised above in this Part.

[312] *PCAST Report* (n 99) 57, 68. See also Jonathan J Koehler, 'Proficiency Tests to Estimate Error Rates in the Forensic Sciences' (2013) 12(1) *Law, Probability and Risk* 89, 95–6.

[313] Where we do not *know* about performance, that is because the procedure and individuals have not been formally evaluated. See also the recommendations for the reporting of latent fingerprint evidence in the *PCAST Report* (n 99) 96.

[314] Non-empirical estimates, such as those reached through the use of a scale or through opinion restricted to similarities (as discussed in *Tang* (n 9) 688 [30], 703 [83] (Spigelman CJ)), do not circumvent this requirement, for tempering the strength of claims in the absence of evidence about performance is speculative; the opinion may be overstated or irrelevant.

[315] In the absence of performance-based information, decision-makers are obliged to rely on misleading proxies such as demeanour, confidence, experience, formal qualifications, job title and accreditation, along with the legitimacy implicitly conferred by the admission of the evidence in a serious criminal proceeding. PCAST is critical of such factors, instead placing emphasis on the need to formally validate procedures and report error rates: *PCAST Report* (n 99) 6. Relying on other evidence to mediate admission might be reasonable when the person is guilty. Unfortunately, we do not usually know if they are guilty. Reliance on other inculpatory evidence raises problems when the suspect/defendant is not in the images — it is likely to lead observers to mis-identify.

considered and, where appropriate, addressed. Accuracy, limitations and other risks should be proactively disclosed in reports (and in testimony).[316]

With respect to points four and five, we recognise that laboratory studies and test conditions will not always mirror actual uses in investigations and trials. Provided applications do not depart significantly from validation testing, modest extensions with caveats would appear reasonable. Empirical studies provide the appropriate frameworks for evaluating results. Where demonstrable (ie empirically supported) abilities falter or are uncertain, we should not revert to the untested opinions and speculations of purported experts.

Sixthly, to the extent that new types of experts and systems (that combine humans and/or algorithms) are relied upon, courts may need to modify their approaches to expertise and reconsider who might be expected (and permitted) to write reports and testify. Obviously, where the output of an algorithm is relied upon but contested, this will tend to require someone possessing expertise with the algorithm (or the system or similar systems) to address validity and reliability. This might be an expert in information technology, or an engineer, or perhaps a statistician or even a person with expertise in perception and cognition. Where systems are designed to combine the opinions of experts (such as facial examiners and super-recognisers), experts and algorithms, or multiple algorithms, ordinarily persons who understand the system, its limitations and the form of reporting should testify. That is, meta-experts — those with a broad understanding of factors contributing to human and algorithm performance.[317] In most circumstances it will not be helpful to call the humans incorporated into these systems (whether facial examiners or super-recognisers) to testify individually.[318]

Similarly, it may not be desirable to call individual super-recognisers to testify. We may need to develop new ways of securing their opinions, for it is not clear that super-recognisers understand or are capable of articulating the reasons for their decisions about identity.[319] Their opinions may sit awkwardly with rules that require 'specialised knowledge' based on 'training, study or

316 In addition, those using the algorithms should be highly trained in order to avoid the use of inappropriate images and forms of image pre-processing: see Claire Garvie, 'Garbage In, Garbage Out: Face Recognition on Flawed Data', *Georgetown Law Centre on Privacy & Law* (Web Page, 16 May 2019).

317 A useful example of the use of a meta-expert in proceedings where admissibility was contested is the treatment of STRmix — an algorithm designed to address complex DNA samples — in *Tuite* (n 80) 229–34 [115]–[120] (Maxwell ACJ, Redlich and Weinberg JJA).

318 The emergence of procedural problems or irregularities may afford an exception.

319 They may not possess insight: see Palermo et al (n 201) 220.

experience'.[320] Moreover, conventional trial mechanisms, such as relying on cross-examination to identify and convey limitations to triers of fact, may have limited efficacy.[321] How, for example, do you cross-examine an individual about an ability that may operate beneath conscious awareness?[322] Super-recognisers might be willing to proffer explanations but these may or may not have informed their decision and may or may not be consistent with scientific knowledge.

To understand the opinions of a super-recogniser, we should be receptive to the testimony of those who study and possess expertise *on* super-recognisers — ie meta-experts.[323] These experts might explain what super-recognisers are and how their abilities are ascertained, provide insight into studies and limitations, along with performance data on the particular super-recogniser. Similarly, where super-recognisers are incorporated into some kind of system (eg where a 'crowd' of super-recognisers is combined in parallel), if the results of the system are to be explained or challenged, we should expect a meta-expert to explain the design of the system, the way the results are collated and reported, and system performance based on ground truth testing.[324]

Seventhly, courts should begin to think about the kinds of error rates they are willing to tolerate when they receive evidence derived from facial examiners, super-recognisers, algorithms and hybrid systems.[325] In most cases, providing an indication of performance (for a witness, algorithm, or system) will be sufficient to enable a decision-maker to make some assessment of the value of the evidence.[326]

320 *Uniform Evidence Law* (n 6) s 79(1). Consider the common law requirement for there to be membership of a recognised 'field' (of knowledge): see *Tang* (n 9) 713 [142] (Spigelman CJ). Super-recognisers appear to have an ability rather than knowledge. Their opinions are not really based on 'knowledge'.

321 Trial mechanisms have not persuaded courts of the importance of the validity and reliability of scientific, technical and medical evidence.

322 In the same way that it does not make much sense to ask how you can tell the difference between sweet and sour. All you can say is that the experience is different — it is difficult to provide useful information on how you make this comparison.

323 Those are persons who possess knowledge about super-recognisers based on relevant scientific literature and ideally their own studies and publications.

324 See Kemp, Edmond and White (n 3) 27, 29–30.

325 This should not be entirely case-based. Courts should be very careful about using other evidence to mediate the reception of error-prone opinions. Rather, they should be intensely focused on the value of the opinion based on scientific evaluation — validation of the procedure and/or empirical evidence of the individual's ability.

326 For an example involving error rates and fingerprint evidence, see Edmond, 'Fingerprint Evidence in New Zealand's Courts' (n 268) 1, 28.

Where, however, the error rate is high, the costs and risks associated with adducing, contesting and managing opinion evidence at trial may not justify admission.[327] Risks of error are accentuated where opinions have been obtained in ways that are inattentive to cognitive bias.[328] A common example is an ad hoc expert focusing on just one suspect and repeatedly watching the associated images and/or video.[329] This may increase confidence without significantly improving accuracy. Efforts to explain these issues at trial will be time-consuming, resource-intensive and unpredictable.[330]

Error rates are particularly important in cases that are exclusively or primarily contested on identity such that — without additional evidence — some levels of error should make it difficult for rational decision-makers to be satisfied beyond reasonable doubt. In such cases, nontrivial error rates introduce resilient doubts. And, empirically derived performance information — eg 20% error in favourable conditions — should not be overruled by the subjective impressions of error-prone jurors *or* juries *or* judges.[331] Where facial examiners, super-recognisers, algorithms and/or hybrid systems cannot reach a decision or cannot reach a decision without a substantial risk of error, the impressions and credulity of jurors, juries and judges should not be allowed to overcome what is *known* about the frailty of human abilities.

Our trial procedures and system of fact-finding should not be based on misconceptions about human capabilities. Jurors, juries and judges are error-prone with images.[332] In the absence of independent and genuinely probative interpretive abilities, some images and even the occasional case may need to be

[327] *Uniform Evidence Law* (n 6) ss 135(b)–(c), 137. Where POIs are disguised, there may be no reliable means of identification. The fact that a podiatrist (or anatomist, or surgeon, or sports scientist) claims to be able to identify a disguised person by their posture or gait should be dismissed in the absence of empirical evidence of performance: see the English case *Otway* (n 18) [18]–[23] (Pitchford LJ for the Court) and the Canadian case *Aitken* (n 18) 28 [80] (Hall J, Finch CJ and Hinkson J agreeing at 35 [104]). See also the analysis in Emma Cunliffe and Gary Edmond, 'Gaitkeeping in Canada: Mis-Steps in Assessing the Reliability of Expert Testimony' (2013) 92(2) *Canadian Bar Review* 327, 356–8 and Royal Society and Royal Society of Edinburgh, *Forensic Gait Analysis: A Primer for Courts* (Report, November 2017) 6, 25.

[328] See Edmond et al, 'Contextual Bias and Cross-Contamination in the Forensic Sciences' (n 292) 5, 10–11.

[329] Edmond and San Roque (n 51) 22.

[330] They may also be contested by parties, including prosecutors, through appeals to common sense that are inconsistent with the results of mainstream scientific research: see Edmond et al, 'Contextual Bias and Cross-Contamination in the Forensic Sciences' (n 292) 18–20.

[331] Sometimes jurors cannot eliminate reasonable doubt: see, eg, *Pell v The Queen* (2020) 268 CLR 123, 145 [39] (Kiefel CJ, Bell, Gageler, Keane, Nettle, Gordon and Edelman JJ). Although the existence of independent evidence may be significant in evaluating opinions about identity.

[332] See generally Phillips et al, 'Face Recognition Accuracy' (n 153) 6173.

removed from the jury. In some cases, opinion evidence should be excluded because the probative value — that is, the capacity of interpretations of the images to assist with identification — taken at its highest is low, but the risk of unfair prejudice is not.[333] In such cases, the images should not ordinarily be available for purposes of identification.[334] Conversely, in some cases the probative value of interpretations of images will be high and potentially compelling.

Eighthly, courts should be cautious about proprietary claims (and non-transparency) to the extent that they prevent meaningful comprehension of the validity and accuracy of algorithms or other systems, procedures and outputs.[335] If proprietary claims (or complexity) prevent meaningful engagement with and evaluation of evidence — and this includes the ability to identify biases and other limitations — then the evidence should not be adduced and relied upon by prosecutors.[336] If our current laws do not require exclusion in such circumstances then they need to be reconsidered, for our accusatorial system is based on the idea that triers of fact should be placed in a position where they are capable of understanding and rationally evaluating the evidence.[337]

Ninthly, individuals, whether purported experts, super-recognisers, facial examiners, or jurors and judges, are all vulnerable to a range of unconscious

[333] Indeed, there are real dangers in allowing potentially weak opinion evidence to contaminate criminal proceedings: see *Volpe* (n 115) [70], [73]–[74], [78] (Priest, T Forrest and Weinberg JJA).

[334] Images may be relevant for other reasons, such as where actions are contested — was a weapon carried? Who threw the first punch? Or, the defence might want the jury to consider the low quality of images if identification evidence is admitted. Where images are presented for reasons other than identification there may be a need to address the risk that the jury will undertake its own comparison even if instructed not to. In some circumstances, it may be possible to excise or pixelate to prevent the jury engaging in comparisons. Courts should not underestimate the risks arising from showing images, or the limits of directions and warnings to influence human cognition and decision-making: see above Part II(B).

[335] See the discussion in *Tang* (n 9) of an anatomist who was unwilling to disclose her procedure for proprietary reasons: at 709–10 [112]–[127] (Spigelman CJ), citing *Re BLM* (District Court of New South Wales, Blanch DCJ, 14 September 2005).

[336] We accept that this is not standard practice with respect to technologies, but respecting secrecy in the absence of independent evaluation raises serious questions about the value of procedures and technologies and the ability to understand and contest them. Consider Rebecca Wexler, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System' (2018) 70(5) *Stanford Law Review* 1343, 1346–56.

[337] Juries should, in principle, be placed in a position where they might evaluate contested expert opinion evidence. They should not be obliged to simply defer to those presented by the state as experts or to rely on misleading proxies such as experience, confidence, and demeanour: see Ronald J Allen and Joseph S Miller, 'The Common Law Theory of Experts: Deference or Education?' (1993) 87(4) *Northwestern University Law Review* 1131, 1133. Cf Brewer (n 79) 1538–9.

influences that bias perceptions and interpretations.[338] These risks are more severe, more pervasive and much more resistant to instructions and conscious effort than courts have recognised.[339] Risks created by suggestion, for example, are one of the main reasons biomedical researchers employ double-blind clinical trials,[340] for, notwithstanding their lengthy medical training and experience, biomedical researchers recognise that they are incapable of overcoming the range of subtle and sometimes unconscious factors and cues that influence interpretation and behaviours.[341] They design research protocols accordingly.[342] Courts should be concerned about the manner in which facial comparisons are undertaken, what the examiner knew and was shown, and what attempts were made to avoid some of the most obvious dangers.[343] (Most comparisons by decision-makers are also vulnerable to suggestion and other forms of unconscious bias).[344]

Mainstream forensic sciences are beginning to respond to the risks of cognitive bias.[345] Modifications such as blinding, sequential unmasking, and full disclosure of the information provided to the expert should be used to reduce errors in image interpretation.[346] It is important to stress that purported experts (as well as juries and judges) are not usually attentive to, and so not shielded from, notorious dangers. Asking jurors to identify the person in images during

338 See, eg, Robertson et al, 'Super-Recognisers Show an Advantage for Other Race Face Identification' (n 198) 213.

339 See Dror, Charlton and Péron (n 278) 75–7; Gary Edmond and Kristy A Martire, 'Just Cognition: Scientific Research on Bias and Some Implications for Legal Procedure and Decision-Making' (2019) 82(4) *Modern Law Review* 633, 633–4 ('Just Cognition').

340 Edmond and Martire, 'Just Cognition' (n 339) 650, 660.

341 Ibid 660.

342 Ibid.

343 Frequently, requests to have purported experts compare images of a robbery with images of a suspect are highly suggestive but inattentive to the risks raised by suggestion and information that is not required to compare images. In many cases requests are supplemented by the provision of domain-irrelevant information, such as the nature of the crime, beliefs of investigators, suspects' criminal records or other incriminating evidence, even though this is not required but likely to mislead: see, eg, *Tang* (n 9) 685 [17] (Spigelman CJ); *Honeysett (NSWCCA)* (n 93) 156 [16], [18]–[23] (Macfarlan JA).

344 Edmond and Martire, 'Just Cognition' (n 339) 633–4, 636.

345 See, eg, Robertson et al, 'Super-Recognisers Show an Advantage for Other Race Face Identification' (n 198) 213–14.

346 See, eg, Dan E Krane et al, 'Sequential Unmasking: A Means of Minimizing Observer Effects in Forensic DNA Interpretation' (2008) 53(4) *Journal of Forensic Sciences* 1006, 1006. Sequential unmasking is a modified form of blinding where the analyst may require, or benefit from, exposure to additional information. The method facilitates the exposure to this additional information in a linear sequence, but requires the documentation of (provisional) results along the way. It is a method intended to assist with the management of contextual bias.

a trial is not a neutral task. Rather, it is highly suggestive. The person *said to be in the images* is sitting suggestively in the dock.[347] Defendants are often selected, in part, because they resemble the person in the images.[348] Suggestive procedures are not conducive to accurate interpretations, and are likely to be most detrimental where the defendant is not actually the POI in the images. These are not trivial risks; research confirms that where the defendant is not the POI, it is *likely* that suggestive evidence or procedures will generate misidentifications.[349]

Facial recognition systems are also vulnerable to bias.[350] Some of these algorithms may have been biased by assumptions, the programming or the sets of faces used to train them.[351] These are potentially significant issues, but we should not overlook the fact that most algorithms have been tested in conditions where ground truth was known. Undoubtedly, algorithms introduce risks, but some of the risks from bias can be understood, circumvented or managed in ways that are much more difficult to address in the realm of human comparison. Moreover, the abilities of all purported experts, unlike those of most algorithms, are simply unknown. We should not assume that algorithms are more biased or more error-prone than humans.[352] Relying on jurors is unlikely to overcome bias or improve accuracy.

Finally, scientific research should underpin and inform legal rules, procedures and jurisprudence. Trial and appellate courts have rarely been presented with relevant scientific knowledge. The leading cases afford remarkably little

[347] In some ways this resembles an in-court, or dock, identification.

[348] In many, perhaps most, cases there is other evidence linking the defendant to the criminal act. In *Tang* (n 9), the defendant was implicated by other offenders as an accomplice, and his fingerprints were matched with those found on stolen cigarettes: at 683 [4] (Spigelman CJ). In *Honeysett* (n 8), police recovered a DNA profile matching the accused: at 127 [8] (French CJ, Kiefel, Bell, Gageler and Keane JJ).

[349] See Helen M Paterson and Richard I Kemp, 'Comparing Methods of Encountering Post-Event Information: The Power of Co-Witness Suggestion' (2006) 20(8) *Applied Cognitive Psychology* 1083, 1098.

[350] See, eg, Phillips et al, 'An Other-Race Effect' (n 193) 10. By not attending to issues of validity and reliability, or failing to consider the alternatives currently in use, some reports and public discussion might be misconceived. See also Australian Human Rights Commission, *Human Rights and Technology* (Discussion Paper, December 2019) 84.

[351] They might also be influenced by procedures and hardware. One recent example concerns variation in the accuracy of an algorithm at male and female face matching: see Vítor Albiero et al, 'Analysis of Gender Inequality in Face Recognition Accuracy' (Conference Paper, IEEE Winter Applications and Computer Vision Workshops, 2020) 85–6. The problem was partly corrected by controlling reference images by the angle of each subject's head pose. Cameras were unwittingly set at a height better suited to capturing the facial images of men who are on average taller; hence the algorithms performed better with the more informative images.

[352] See above Figure 7.

evidence that critical issues with unfamiliar face matching have been identified, understood, incorporated into the interpretation of legal rules and procedures, or effectively conveyed to decision-makers during criminal proceedings.[353] Admitting speculative opinions and leaving their evaluation to an assessment of weight has placed seemingly unbearable strains on criminal procedure and personnel. There is *no evidence* that cross-examination or judicial directions and warnings are effective at exposing limitations in ways that improve decision-makers' assessment of opinions about identity or their interpretations of images.[354] This should be contrasted with the extensive scientific research confirming the susceptibility of ordinary humans to error in identifying unfamiliar faces, a vulnerability accentuated when the opinions of those presented as experts are themselves wrong.[355]

V Conclusion

This article presents an evidence-based case for reconsidering our traditional approach to identification evidence. All identification evidence is interpretive — ie opinion — and so we should redesign our rules and jurisprudence to reflect that reality. Any revision should maintain scope for sensory witnesses (eg eyewitnesses) to express opinions about what they directly saw, heard or otherwise perceived. When it comes to the identification of persons in images and video recordings, opinions should be regulated with close attention to the abilities of the individuals testifying — and this should be informed by mainstream scientific research and an awareness of the risks. Apart from eyewitness evidence, opinions about identity should not be admissible unless there is evidence that the 'witness' is familiar with the person they purport to identify from interactions beyond the criminal investigation or is demonstrably expert at the identification of persons in images.[356] Abilities must be supported by scientific knowledge so that the opinions are probative and susceptible of rational evaluation. Now that we have access to genuine experts, we should not rely on the speculative opinions of purported experts whose abilities are unknown or whose perceptions have been irreparably contaminated by their role in the

[353] See the discussion above in Part II(A) of *Tang* (n 9), *Honeysett* (n 8) and *Smith* (n 20).

[354] The kinds of directions, instructions and warnings offered to juries are derived from the collective wisdom of judges with some limited input from social sciences. They appear to have very limited utility, in relation to images, other than to indicate the need for caution.

[355] See above Part III(B)(1).

[356] The use of prison familiars and prison officers should probably be avoided. Our trial mechanisms provide few means of genuinely scrutinising such opinions and they tend to raise issues of unfair prejudice. We also have access to alternative independent methods.

investigation. We should not trust juries to recognise or somehow transcend interpretative difficulties that are notorious among scientific specialists. Such expectations are unrealistic and unnecessary.

The need to engage with scientific research on validity and reliability is not restricted to the use of images and identifying POIs in images. It applies to the identification of persons by voice, as well as most of the other feature comparison procedures — such as fingerprints, firearms and tool marks, shoe and tyre marks, blood spatter, documents, fibres, glass, soil, drugs, paints, gait and so on.[357] When admitting and reviewing forensic science and medicine evidence, courts should routinely ask about validity, reliability and proficiency. That is: can you do the specific task? How often do you make errors? And how do we *know*? Where answers are not based on formal scientific evaluation, courts should be uneasy. Judges must be willing and able to exclude evidence that is unreliable or not known to be reliable, for speculative opinions are difficult to reconcile with the legal requirement of knowledge. In our rational tradition, they threaten the goals of rectitude, fairness and efficiency.

[357] See the discussion in Edmond, 'What Lawyers Should Know' (n 9) 33, 41.