Preventing machines from lying: why interdisciplinary collaboration is essential for understanding artefactual or artefactually dependent expert evidence

Tim J Wilson, [corresponding author] Law School, Northumbria University  
(tim.wilson@nothmbria.ac.uk) Jesper Bergman, Department of Computer and Systems Sciences, Stockholm University, Adam Jackson, Law School, Northumbria University and Oliver B Popov, Department of Computer and Systems Sciences, Stockholm University

Abstract

This article demonstrates a significantly different approach to managing probative risks arising from the complex and fast changing relationship between law and computer science. Law’s historical problem in adapting to scientific and technologically dependent evidence production is seen less as a socio-techno issue than an ethical failure within criminal justice and state institutions. This often arises because of an acceptance of epistemological incomprehension between lawyers and scientists. Something compounded by the political economy of criminal justice and safeguard evasion within state institutions. What is required is an exceptionally broad interdisciplinary collaboration to enable criminal justice decision-makers to understand and manage the risk of further ethical failure. If academic studies of law and technology are to address practitioner concerns, it may often be necessary, however, to step down the doctrinal analysis to a specific jurisdictional level.

Key words

Explaining/understating AI/ML-assisted decisions; interdisciplinary methodology in law and technology studies; neoliberalism; ethics and criminal justice systems

Introduction

Of course, machines cannot lie any more than, as Alan Turing observed, nothing could be gained by asking whether they could think. Hence, the term ‘Artificial Intelligence’ (AI). Turing reformulated the latter question as whether human interrogators ‘could be taken in by cunningly designed but quite unintelligent programs’? Likewise, expert evidence that is unreliable because of error in or misunderstanding about data processing by computer programmes trained by Machine Learning (ML), hereafter ‘artefactual/artefactually dependent evidence’, clearly cannot be termed lies. Allowing verdicts or sentencing decisions to turn on unsound, potentially unsound or misunderstood artefactual/artefactually dependent evidence or being complicit in the evasion of probative safeguards, however, is normatively equivalent to negligently or knowingly colluding with perjury. This analogy, reflecting the ‘legal, moral and social foundations of criminal law’, highlights the importance of ensuring that the trial of fact must not be invalidated because of avoidable errors or

misunderstandings in AI-assisted decision-making. The responsibility to get this right applies to all individuals who, as expert witnesses, investigators, lawyers or factfinders use such evidence. Those who should have the expertise to prevent non-expert decision-makers being misled by or misunderstanding opinion evidence, however, bear the greatest professional responsibility in this respect. A similar ethos prevails in evidence-based medicine, where the correlation (not causality as in criminal law) of data - the driving force of AI/ML processing – is normally sufficient, but best practice still mandates critical assessment of artefactual outputs by professional decision-makers.\(^4\) Doctors must understand the limitations of an AI/ML tool before using it.\(^5\)

This article conjoins socio-legal research in England and Wales with computer science research in Sweden. Both were part of a research project with other international partners into the problems of police detective work on the TOR-protocol, an anonymous communication network (ACN) with hidden services that can be used for digital marketplaces etc. as part of the Dark Web.\(^6\) The English research had earlier resulted in a paper that looked at aspects – personnel, organisational, cultural and ethical issues - of functional adaptation by the police in response to crime involving digital communication and services.\(^7\) The Swedish contribution is based on ‘proof of concept’ research into the design of an AI/ML-encoding tool to improve the effectiveness and forensic soundness of dark web cybercrime investigations.\(^8\) This article goes wider than police dark web investigations because the English co-authors have also drawn on insights from earlier research into forensic DNA and fingerprint comparisons. Our interdisciplinary co-authorship was essential for insight into what is required to adapt to increasing reliance on AI-assisted decisions in criminal justice. In which respect, ‘understanding the strengths and limitations of methods employed in the CJS is essential, in order that investigators and courts know what may not have been found or what artefacts may be present’\(^9\).

The article is written on the assumption that readers will be knowledgeable in the law of evidence, but not necessarily AI/ML. What may at first sight seem to be an abstruse subject, is rooted in professional experience by addressing two questions faced by prosecutors when the facts of the case mean that decision-makers could not make a properly informed decision without assessing the reliability/weight of evidence of artefactual or artefactually dependent evidence: -

\(^4\) The US Agency for Healthcare Research and Quality (AHRQ) definition of evidence-based practice: ‘A way of providing health care that is guided by a thoughtful integration of the best available scientific knowledge with clinical expertise. This approach allows the practitioner to critically assess research data, clinical guidelines, and other information resources in order to correctly identify the clinical problem, apply the most high-quality intervention, and re-evaluate the outcome for future improvement.’ <https://www.ahrq.gov/topics/evidence-based-practice.html> on 25 September 2022 (emphasis added).


\(^6\) Other colleagues were affiliated to the Dutch Open University, the Amsterdam University of Applied Sciences (HvA), the NHL Stenden University of Applied Sciences, the Dutch Police Academy and the Politihøgskolen (Norwegian Police University College).


1. Is a digital specialist’s testimony expert witness evidence or not?
2. As there are an insufficient number of such specialists trained to be expert witnesses, how can the available experts be prepared or supported as witnesses?²¹⁰

Despite questions being put to the UK coauthors during our related empirical research into how ACNs are investigated in England and Wales, this is an issue that, with relatively nuanced distinctions, will apply in all common law jurisdictions. It is equally likely to be relevant in non-adversarial proceedings and with equal significance for the necessity for expert candour.

Electronic evidence is a large and varied field where there is a risk of mistakenly admitting such evidence as lay testimony about direct evidence when it is frequently indirect evidence.²¹¹ We have confined our answers to both questions to specific circumstances: where artefactual or artefactually dependent expert evidence (i.e., the evidence submitted has relied wholly or in part on a computer programme that searches for statistical correlations or patterns in data). This includes fingerprint and DNA comparisons and extends to automated internet surveillance. Under English law in such circumstances, we argue that expert opinion evidence to be peritus ²¹² must include wider knowledge than just a traditional skill/expertise, for example, in forensic genetics. They should also have knowledge of (i) the AI/ML system/application used in producing evidence that thereby results in AI-assisted decision making and (ii) any relevant interdisciplinary insights that, as explained later, may include social as well as STEM sciences. As a result of focusing on the duty of those producing artefactual or artefactually dependent evidence to ensure that their reliance on AI/ML and any consequences/risks arising from this is always understood by anyone making such AI/ML assisted decisions, it will be seen that the response to the second question overlaps epistemologically with the first.

Guidance published in 2022 by the Information Commissioner’s Office (ICO) and the Alan Turing Institute (Turing) about explaining the use of AI/ML in decision making (hereafter the ‘explainer approach’) enables us to approach this probative issue also from public law and professional good practice perspectives, with analogous examples from medical AI/ML assisted decision-making.¹³ This contemporary perspective together with earlier literature about electronic and artefactual dependence evidence, especially fingerprint comparison and forensic DNA, means that we question a 1990s turn towards a view uncritical of computers as a source of evidence. During the 1980s the potential fallibility of electronic evidence was acknowledged. For example, it was noted that ‘computers must be regarded as imperfect devices’¹⁴ and, where relevant, the court might expect to hear testimony relating to a computer system from its initial development to use in the instant case.¹⁵ Such caution¹⁶ appears to have been put aside – possibly under pressure to follow what was seen

---

¹⁰ During discussion at a PDTOR workshop organised as virtual event by Northumbria University on 14 January 2022 and attended by 38 police/other agency investigators and trainers, CPS prosecutors and an interdisciplinary group of academics from universities in five jurisdictions. It is intended to deposit an anonymised workshop record with the UK Data Service.


¹² R v Silverlock [1984] 2 QB 766, [771]

¹³ ICO and Turing Institute, Explaining decisions made with AI, (ICO/Turing Institute: London, 2022).

¹⁴ R v Minors (Craig); R v Harper (Giselle Gaile) [1989] 1 WLR 441 [443]

¹⁵ R v Wood (Stanley William) [1983] 76 Cr App R 23

¹⁶ Similarly, taking just one example of earlier judicial scepticism, in the 1980s, Californian courts, despite operating a permissive criminal admissibility regime towards expert opinion were generally willing to prevent
as more efficient American and English civil admissibility rules developed from the 1960s\(^{17}\) - and a PACE safeguard in respect of the admission of such evidence was abolished in 1997.\(^{18}\) This consolidated English criminal evidence admissibility doctrine around what by then had come to be seen as a common law rule that computers are ‘reliable’.\(^{19}\) This rule has been enforced by judicial notice and, as will be considered in the second section, sometimes by statute, in all major common law jurisdictions to expedite criminal proceedings by avoiding the need to prove what thereby doctrinally became ‘obvious’ facts.\(^{20}\) This approach, however, was subsequently qualified in 2003 in criminal proceedings in England and Wales, so that where a representation made by a machine (including a computer programme) relies upon information supplied directly or indirectly by a person it must be proven that the information supplied was accurate in order for the evidence to be admissible.\(^{21}\) This article builds on the recognition in the Act of 2003 of the importance of the cognitive link between the human mind and computer processing by highlighting the importance of computer science’s interface with the natural and/or social sciences in the production of artefactually dependent expert evidence.

The 1997 PACE amendment - to implement a Law Commission recommendation - was undertaken with rare alacrity (almost on publication) and justified solely or primarily by Tapper’s almost passing remark that ‘most computer error is either immediately detectable or results from error in the data entered into the machine’.\(^{22}\) Mason noted that the Commission ignored ‘a great deal of technical material in the 1970s and 1980s [demonstrating] that software errors might not be obvious’.\(^{23}\) He also endorsed Ormerod’s argument at the time, based on Dillon, that where digital evidence is fundamental, the prosecution is not entitled to rely on a presumption to establish facts central to an offence.\(^{24}\) In addition to looking more closely at Tapper’s caveat about the scope ‘for error in the data entered into the machine’, we also draw on a 2020 article solicited by the Law Commission to review the basis for the 1997 amendment\(^ {25}\). This authoritative deployment of scientific scepticism against the justification for the safeguard’s abolition also provides valuable insights about the scope for malfunction within the technological aspects of artefactually or artefactually dependent expert evidence production.

\(^{17}\) The civil developments were summarised in C. Tapper, ‘Discovery in Modern Times: A Voyage around the Common Law World’ (1991) 67 Chicago-Kent Law Review 217

\(^{18}\) The Youth Justice and Criminal Evidence Act 1999, ss. 60, 67(3), Sch. 6 repealed The Police and Criminal Evidence Act 1984 s69 (1) rule that computer evidence was only admissible if certified to have been working correctly.

\(^{19}\) ‘In the absence of evidence to the contrary, the courts will presume that mechanical instruments were in order at the material time’. S. Mason ‘The presumption that computers are “reliable”’ in Mason and Seng (eds.), above n. 19 at 101

\(^{20}\) Ibid. at 104-107

\(^{21}\) Criminal Justice Act 2003, c.44, s.129

\(^{22}\) Law Commission, Evidence in Criminal Proceedings: Hearsay and Related Topics LC245 91997) para. 13.7; Above n.18 at 248; The Commission failed to acknowledge that article’s subject was efficiency in English and American civil litigation and in the paragraph concerned criticised a ‘quite extraordinarily lax’ English statutory admissibility rule

\(^{23}\) See Mason, above n. 19 at 171-173


A change of approach in legal proceedings is not required, however, to reverse undue deference to evidence production involving computers. A rigorously scientific approach to the use of Part 7 of the Criminal Practice Direction (Crim PD) would be sufficient. To be admissible in criminal proceedings in England and Wales any expert opinion evidence must be

(a) ‘sufficiently reliable’. (Crim PD, 7.1.1(d)): the court is empowered to make a pre-trial determination of the reliability of such evidence which includes consideration of the validity of any methodology employed by the expert (Crim PD 7.1.2(b)) and whether the expert’s methods followed established practice in the field (Crim PD 7.1.2(i)); and, similarly,

(b) reliant ‘on an examination, technique, method or process which was not properly carried out or applied, or was not appropriate for use in the particular case’ will be indicative of a lack of reliability (Crim PD 7.1.3(d)).

What is easily stated as doctrine, however, as we show, is not necessarily readily achievable in practice.

There is limited evidence to suggest that such challenges are being brought in practice and, as is noted below, despite various admissibility reforms unscientific or problematic evidence generally faces weak scrutiny. Hence, our argument that the identification of problems with artefactual or artefactually dependent expert evidence technical systems is reliant either on both the knowledge of the expert that such problems exist and the expert’s candour in revealing them. We explain why we are sceptical about the ability of defence experts or the court to independently identify and resolve potential problems. Hence, the thrust of this article is the need for clear and comprehensive expert candour.

We begin with an explanation of the interdisciplinary knowledge gap, and how this overlaps and interconnects with an institutional tendency to safeguards evasion and the impoverished political economy of criminal justice. The second section considers how the knowledge gap has been exemplified in formal procedural rules, caselaw and scholarship dealing with admissibility and the ultimate issue rule. This leads to the proposal (after noting highly relevant analogies with medical best practice) that for an expert witness to be effectively peritus they should become- in ICO and Turing terminology - ‘explainers of risk in AI-assisted decisions’. The next section begins by analysing cultural inhibitors to narrowing the interdisciplinary knowledge gap, and then illustrates the critical importance of interdisciplinary knowledge in computer science CJS focused research, development and use. The fourth section brings together the two themes – evidence in court and end-user relevant research – that emerge from the analysis. It suggests that being able to explain AI-assisted decisions or evidence production results in the type of professional insight needed to inform the development, operationalisation and upgrade life-cycle stages of AI/ML applications. Such coproduction is essential for significantly and systemically reducing reliability risks in artefactual or artefactually dependent expert evidence.

The interdisciplinary knowledge gap, a tendency towards safeguards evasion and economic (organisational and ideological as well as quantum) influences

Unlike the speculative questions in Alan Turing’s paper, problems and risks arising from poor AI-assisted decision-making in criminal justice are not theoretical. There is a long and shameful history of courts globally being ‘taken in’ by expert evidence based on what is sometimes, not

inappropriately, referred to as ‘junk’ science’. Misrepresentation of or misunderstandings about genuine science has also resulted in major miscarriages of justice. The ‘infallibility’ myth or zero error rate claims of expert fingerprint comparison evidence long survived the first serious attempts to address admissibility systematically and scientifically, with the US Federal Rules of Evidence in 1975 to Daubert in 1993, when US courts effectively grandfathered an inductive fallacy - ‘the uniqueness of all human fingerprints’ - rather than question the accuracy of the identification process. These failures often stem from interdisciplinary knowledge gaps. This risk could be systemic in AI-assisted decisions. For example, intoximeter reliability depends on the accurate application of computer programming, chemistry, and biology. A natural science error can invalidate the results, even if the programming itself is flawless.

Carr et al. have drawn attention to how with Streamlined Forensic Reporting (SFR), ‘assumptions that traditional legal safeguards will identify any weaknesses and strengths in expert evidence misses the valuable opportunity to properly consider the evidence before an admission of guilt may have to be made.’ Irreversible plea or scope of defence decisions - usually made without expert advice - about the validity of scientific evidence predate the court hearing when admissibility safeguards apply. Similarly with electronic/digital evidence generally, a failure to identify a potential defect in such evidence early enough may restrict the ability to challenge the reliability of the evidence. Both are examples of a tendency towards safeguards evasion. The institutional evaluation and justification for the introduction of SFR was devoid of any ‘idea about its impact on the quality of the evidence presented or the rectitude of outcomes’. SFR effectively reintroduced ‘zero error’ fingerprint testimony by a back door that a judge cannot close, having been ‘validated solely against institutional efficiencies and related savings’. Thus, the knowledge gap permits a further drift in safeguards evasion propelled by economic objectives overriding fair trial principles. The risks are likely to be greater with electronic/digital evidence. A SFR report may be submitted by a police officer managing the investigation, who may simply report what he/she believes is ‘useful’

30 Daubert v. Merrell Dow Pharm. Inc.509 U.S. 579
33 S. Carr et al., ‘Demonstrating reliability through transparency: A scientific validity framework to assist scientists and lawyers in criminal proceedings’ (2020) 308 Forensic Science International 110110
34 R v Coultas, [2008] EWCA Crim 3261 [21]
36 Ibid. ‘A finger/palm mark in exhibit MRN/1 is identified to a set of fingerprints held on the National Fingerprint Database in the name of XXXXX reference XXXXX’ at 768; see also 775
37 Ibid. at 789
for their case, and, presumably because of extensive backlogs of work in digital units, without digital specialist endorsement or drawing necessary caveats to decision-makers’ attention.

These three problems have multiple causes that cannot be analysed in a single article, but the economic problems are both endogenous and exogenous to the political economy of criminal justice systems.

Digital investigations and AI/ML dependent evidence production applications - initially, automated fingerprint identification systems (AFIS) – emerged after neoliberalism had become the dominant politico-economic ideology in pluralist democracies. Interdisciplinary knowledge gap risks appear to increase as criminal justice became more reliant on commercially developed equipment. It will be seen in the fourth section how black box/source code challenges are rarely possible in US courts because IPR outweighs fair trial principles. Economic considerations have threatened US professional ethics and probative reliability, with expert testimony vulnerable to contingency fee bias manipulation and “commercial pressure to make proficiency tests easier”.

Contemporary politico-corporate literature places AI/ML devices in the sphere of ‘disruptive technology’ or ‘disruptive innovation, even if modestly stated as improving productivity and efficiency by ‘minimizing administrative and operational overheads within policing’. A example of this wide-spread trend from health care, conceptualises technologically driven innovative as a rapid transition to new models of service provision with more than a hint of workforce deskilling and changes in legal frameworks to profit from the opportunities created by technology to manage ‘rising demand, increasing cost and insufficient funding.”

Similar expectations can be found in

---

38 Concerns reported from a seminar held for lawyers, forensic practitioners, judges and academics at Northumbria University in November 2018, C. McCartney, ‘Streamlined forensic reporting: Rhetoric and reality’ (2019) 1 Forensic Science International: Synergy 83
40 In the UK by 1979: C. Crouch, The Strange Non-Death of Neoliberalism (Polity: Cambridge, 2011), 20-23
43 Deloitte & Touche (Middle East), Digital Policing Disruptive technologies enhancing the policing experience (Deloitte & Touche (ME): Dubai, 2020), 8 and 10. If Deloitte’s literature is typical, the unmissable ‘disruptive’ emphasis now appears to have switched to the Middle East. The Changing Landscape of Disruptive Technologies: Tech hubs forging new paths to outpace the competition, (KPMG International Cooperative: Zug 2018) - then in its sixth (possibly final) year – is an earlier example of a consulting firm popularising managerial literature for European politico-corporate consumption.
46 Expert Panel on Effective Ways of Investing in Health (EXPH), Disruptive Innovation - Considerations for health and health care in Europe (European Commission: Brussels, 2016) 7 and 13. Arguably, with more extreme examples, disruption theory has been captured to justify often unsustainable market practices under the guise of ‘innovative’ business models. See: Mark Findlay et al., AI and Big Data: Disruptive Regulation (Edward Elgar: Cheltenham, 2023)
unexpected contexts, for example, by ‘reimagining the human role and contribution within the evolving human–machine cognitive system’ for military decision-making, where the role of military commander needs to evolve from controller to teammate. In this brave new world it could prove difficult to show where command responsibility for war crimes might lie.

A more immediate question for criminal justice professionals and policy makers, however, arises from whether research into the impact of AI/ML by focusing on manufacturing and services industries, has generally failed to examine its impact on knowledge-intensive activities. Ribeiro et al. suggest - at least in biosciences - that ‘routine tasks do not necessarily disappear … and challenge the assumption that automation and digitalisation contribute to productivity in exclusively positive ways.’ While we cannot comment on whether this ‘digitalisation paradox’ also applies to criminal justice, the arguments in this article support the case for similar research into the automation of criminal justice knowledge-intensive work.

This is not to deny the scale of the challenge created for criminal justice professionals and governmental budgets from the volume and complexity of crime arising from the digitalisation of everyday life. For example, the industrialised scale and organisation of cybercrime can be seen from initial reports of a single international dark web police operation. The iSpoof takedown, identified 59,000 potential suspects (with an estimated 200,000 victims in the UK alone) who purchased access to its cyber-fraud enabling services. The use of the term ‘Dark Web’ for anonymous communication networks and services adds to the complexity of criminal justice responses. The TOR-protocol has many legitimate uses, such as protecting journalists’ sources, whistleblowers and access to uncensored information. Dealing with rising demand in such sensitive and complex contexts requires the exercise of human discretion and judgement, even if such decision making can be usefully supported by automation. T

The approach adopted by the applicable CrimPD in England and Wales, provides an alternative to rigid presumptions about reliability or accuracy of digital processing. The Directions provide scope for all types of expert evidence. to be tested, pre-trial, against a series of factors to determine evidentiary reliability. Whilst doctrinally this should avoid rigid presumptions about any form of expert evidence, discussed briefly above, there is limited evidence to support the fact that, in practice, such challenges are routinely being brought. This article by analysing the wide and varied scope for error in artefactual or artifactually dependent evidence, indicates that with such expert evidence, there may be substantial risks of erroneous presumptions about reliability.

Guidance about making processes, services and decisions delivered by AI/ML intelligible by default has emerged from within the computer science community, hence, the ICO and Turing ‘explainer’ approach that is applied in the fourth section to expert artefactual or artefactually dependant evidence. First, however, the next two sections analyse why this approach is potentially so valuable in criminal justice: respectively, because of the problems encountered in successfully adapting

---

48 Ribeiro et al., ‘The digitalisation paradox of everyday scientific labour: How mundane knowledge work is amplified and diversified in the biosciences’ (2023) 52 Research Policy 104607
50 Tom Symonds, ‘Police text 70,000 victims in UK’s biggest anti-fraud operation’, <Police text 70,000 victims in UK’s biggest anti-fraud operation - BBC News> on 24 November 2022
51 Since 2019 the BBC News website and its programmes have been accessible via the Tor-protocol: <BBC News launches ‘dark web’ Tor mirror - BBC News> on 24 November 2022
52 Crim PD, 7.1.2-3
criminal justice systems to scientific developments (including computer science) to improve the quality of criminal justice; and the critical importance of interdisciplinary collaboration for ensuring successful adaptation or avoiding serious epistemological error.

**Expert witness artefactual or artefactually dependant evidence: admissibility and the ultimate issue rule**

The first question - is a digital specialist’s testimony expert witness evidence or not - might surprise some readers. It partly reflects a cultural legacy from the beginning of digital investigations/ digital forensics in the 1980s. Basic evidential considerations were soon adhered to, as can be seen with the emphasis on chain of custody requirements in most descriptions of ‘forensic soundness’ in investigative work that relied on computers. Insufficient consideration, however, was given to evaluation of the results or alternative interpretations for such results\(^{53}\) or even whether digital artefacts recovered and analysed during a digital forensic investigation might have been tampered with before seizure.\(^{54}\) The idea of digital work as a technical operation rather than a scientific practice may have been reinforced by the marketing of ‘effectively idiot-proofed’ applications. Where such mindsets still prevail, ‘digital forensic practitioners incorrectly assume that they are simply reporting what they observe and are unconscious of the interpretations and decisions inherent in digital investigations.’\(^ {55}\)

In England and Wales such naïve confidence is likely to have been reinforced by the turn in legal and governmental thinking with consolidation around the common law presumption of computer reliability. This would have been amplified by a statutory requirement in many common law jurisdictions that results from government approved AI/ML devices (e.g., intoxilisers and genotyping software) must be treated as accepted fact. In England and Wales intoxilisers approved under statutory powers are erroneously (see fourth section) presumed to be reliable even following untested modifications.\(^ {56}\) The reliability of these systems cannot be readily challenged.\(^ {57}\)

The admissibility of expert opinion evidence in most major common law jurisdictions\(^ {58}\) hinges on relevance and reliability.\(^ {59}\) The former results in the doctrinal requirement that expert opinion evidence is only admissible where it provides factfinders with ‘scientific information which is likely to be outside the experience and knowledge of a judge or jury.’\(^ {60}\) For this the witness, as discussed above, must be competent or ‘peritus’\(^ {61}\)

A general lack of statutory definitions/specifications for expert qualification and training provides opportunities for a range of errors and omissions, particularly the admission of opinion evidence by

---

\(^{53}\) Whitcomb, above n. 39

\(^{54}\) C. Neale et al. ‘The case for Zero Trust Digital Forensics’ (2022) 40 *Forensic Science International: Digital Investigation* 301352


\(^{56}\) For example, s. 7(1)(a) of the Road Traffic Act 1988

\(^{57}\) See Mason, above n. 19 at 177-179

\(^{58}\) G. Edmond et al., ‘Admissibility Compared: The Reception of Incriminating Expert Evidence (i.e., Forensic Science) in Four Adversarial Jurisdictions’ (2013) 3 *University of Denver Criminal Law Review* 31-109


\(^{60}\) R v Turner [1975] QB 834 [841]; now codified in Criminal Practice Direction (CPD) 19A.1

\(^{61}\) R v Silverlock [1894] 2 QB 766, [771]
individuals who should not be treated as experts. 62 English and American courts appear to have glossed over the epistemological distinction between lay testimony of fact about reliance on or the use of a computer compared with expert scientific evidence about its reliability. Though, as Mason observes, there may often be a fine dividing line between lay evidence about the day-to-day operation of a system and expertise in its operation, and expert opinion about the operation of computer systems. 63

Within many common law jurisdictions, admissibility safeguards are applied with what US judges have described as a generally ‘liberal’ or ‘permissive’ approaches, hence, admissible evidence might be ‘shaky’. 64 Comparative analysis suggests that ‘admissibility standards have not contributed to the exclusion (or informed systematic evaluation) of unreliable and speculative forms of incriminating opinion evidence in courts.’ 65 English practitioner experience is that ‘the working principle of assumed reliability appears to be the default position’. 66

Admissibility decisions about AI/ML dependent evidence may turn on an exceptionally fine line. For example, the Court of Appeal in Dlugosz 67 endorsed the admissibility of expert statements initially based on AI/ML application results that were subsequently criticised by fellow scientists for substantial professional reasons: the testimony exceeded what reliable methodology in the scientific literature, training or standards would allow. 68

An alternative view on Dlugosz, however, suggested by Ward, and based on analogy with hearsay evidence (must be ‘potentially safely reliable’ in the context of the evidence as a whole) is to recognise the potential value of ‘expert evidence of weak or unknown probative value ... adduced as one part of a body of evidence which taken together is arguably compelling’. 69 He sees indications in the judgment that the Dlugosz DNA evidence was seen ‘as quite close to the borderline’. 70 Non-scientific evidence discovered because of the AI/ML DNA outputs probably convinced the Court of Appeal that justice had been done for the victim. Forensic Science Regulator (FSR) guidance subsequently issued in response to this case did not take an exclusionary stance. It confirmed that the evidence had been presented in a scientifically erroneous manner and advised how the results obtained from the AI/ML search application should have been presented more neutrally and with frankness about their weak probative value. 71

---

63 See Mason, above n. 19 at 341-349
65 See Edmond et al., above n. 58 at 108
66 See Carr et al., above n.33 is a particularly authoritative assessment of risks within England and Wales, being co-authored by legal/scientific academics with legal practitioner and forensic science experience, including Angela Gallop, the doyen of UK forensic science.
67 R v Dlugosz, Pickering and MDS (2013) EWCA Crim 2
68 C. Champod, DNA transfer: informed judgment or mere guess work? (2013) 4 Frontiers in Genetics Article 300
70 Ibid. at 248
71 Forensic Science Regulator, FSR-G-222: DNA Mixture Interpretation Issue 3 (FSR: Birmingham, 2020), Guideline 12 para. 7.2.2
What we term the interdisciplinary knowledge gap prompted Ward to ask whether it is simply unrealistic to expect prosecution and defence lawyers, judges or juries to detect, unaided, ‘the ways in which an expert’s necessarily simplified account of the science unduly favours one party’? One option is for judges and lawyers to keep abreast of scientific developments through the work of ‘key epistemic “monitors”’, such as the FSR and the US National Academies of Science (NAS), and ‘if possible, to be informed of any cogent criticisms of those bodies’ work.’ This begs the question, however, of whether busy lawyers and, we suggest, also investigators, can be expected to keep abreast of, in the case of the FSR, voluminous guidance that is highly technical, subject to regular revision and written essentially for the relevant expert scientific communities. For historical reasons, the FSR guidance is generally a remedial and – even in its 2023 statutory incarnation – is still an incomplete response to known problems and progress in extending its coverage is necessarily slow. Evidence that is not Code-compliant remains admissible (and the weight to be attached to it) remains a matter for case-by-case decisions. The judicial gatekeeping role becomes even more difficult when admissibility turns on the significant minutiae of computer science. While one American judge learned to code in Java in preparation for the copyright dispute, ‘do most judges even possess the technical knowledge to understand coding languages?’

In a later paper Ward canvasses another option for all forensic science testimony identical to the explainer approach: the professional ethics and legal duties of expert witnesses should require the revelation of uncertainties – ‘where there are possibilities of error, bias, disagreement or alternative explanation’ - to assist CJS decision making.

The explainer approach is consistent with the ultimate issue rule: even when a decision turns on a matter which the tribunal would be unable to understand ‘without the assistance of experts’, ‘the power of decision is retained by the tribunal of fact’ Expert witnesses should be careful to recognise ‘the need to avoid supplanting the court’s role as the ultimate decision-maker on matters that are central to the outcome of the case’. Commentators have noted how the significance of this rule has been diminished, or dismissed as ‘a matter of form rather than of substance’. Strict compliance with the rule may certainly be undesirable in certain circumstances, such as diminished responsibility cases, where the clinical symptoms diagnosed by the expert are used to explain the events. Hence, the jury were discouraged in Golds – a case involving expert evidence unchallenged by the prosecution - from making themselves ‘amateur psychiatrists’. Whatever the current status and detailed application or definition of the rule itself, there remains considerable authority for the

---

72 Ward, above, n. 69 at 249.
73 Ibid. at 250.
77 Davie v Edinburgh Magistrates [1953] SC [34]
78 Phipson on Evidence 20th edn. (Sweet & Maxwell: London) 33-12
79 Pora v The Queen [2015] UKPC [9].
80 Roberts and Zuckerman, above n. 3 at 490
81 R v Stockwell (1993) 97 Cr. App. R. [260] [265]; see also: R v Atkins [2009] EWCA Crim 970
82 See Ward, above n.69 at 243
83 R v Golds [2016] UKSC 61 [51]
84 R v Robb (1991) 93 CRApR [161] [165]; R v Luttrell & Others [2004] EWCA Crim [1344]
view that the evaluation of the reliability of an expert evidence remains the role of the tribunal of fact aligned – consistent with supported by doctrinal analysis that separates the expert and decision maker’s roles\textsuperscript{85} - with a warning that experts must not trespass upon jurisprudential territory and confine themselves ‘to purely scientific questions, leaving open any issue as to the surrounding facts’.\textsuperscript{86} Otherwise - as Biedermann and Kotsoglou have commented - with the court’s complicity, an expert witness would usurp the factfinders’ normative role, for example, in making legally significant judgements with the risk, for example, of incorrect false identification and, hence, false incrimination of a defendant.\textsuperscript{87}

In England and Wales, moreover, the Rule’s function has been authoritatively preserved in guidance about how judges should deal under Part 7 of the CrimPD, with any issues relating to the reliability of expert evidence raised pre-trial. This will form part of the judge’s determination regarding the admissibility of the evidence. Where the evidence is sufficiently reliable to be admitted, any dispute as to the reliability of the evidence will be addressed in open court to assist the factfinder in judging the weight to be attached to the evidence. The Crown Court Compendium\textsuperscript{88} offers guidance to judges on the direction to be given to juries, including in the following terms:

“...as with any other witness, it is the jury’s task to weigh up the evidence of the expert(s), which includes any evidence of opinion, and to decide what they accept and which they do not... Any factors capable of undermining the reliability of the expert opinion or detracting from his/her credibility or impartiality should be summarised. The reliability factors listed in CrimPD Ch 7 reflect the common law, and should be used to assist the jury in evaluating and assessing the weight of the expert evidence. It may be that not all these factors will be under consideration during the evidence and therefore the direction and the factors should be tailored to the issues in the case.”\textsuperscript{89}

As a result, the experts themselves (both prosecution and defence) have a key role in identifying issues of evidentiary reliability and in assisting the court to understand them consistent with the expert’s overriding duty as an “objective and unbiased”\textsuperscript{90} assistant to the court.

Expert witnesses – when diligently seeking to fulfil this assistive role - still need to overcome the problem noted by Bollé et al. that ‘[m]any existing ML approaches lack sufficient transparency and reproducibility for forensic purposes, and are not designed in a way that helps forensic practitioners evaluate and explain the outputs of automated systems effectively’.\textsuperscript{91} The next section analyses the cultural and institutional inhibitors to overcoming the interdisciplinary knowledge gap, and, more generally, how the lack of interdisciplinary collaboration may compromise the reliability of evidence

\footnotesize{\textsuperscript{85} ‘As a general rule, forensic scientists and other expert witnesses should be assessing the probability of the evidence, rather than commenting on the probability of contested facts (much less the ultimate issue of guilt or innocence).’, C Aitken et al., Fundamentals of Probability and Statistical Evidence in Criminal Proceedings (RSS: London, 2010), 112}

\footnotesize{\textsuperscript{86} R v Jones (William Francis) [2020] EWCA Crim 1021 [36] and [38]}


\footnotesize{\textsuperscript{88} Judicial College, The Crown Court Compendium: Part I: Jury and Trial Management and Summing Up, June 2023, Ch.10-3 para 18(2)(b)\textsuperscript{89}}

\footnotesize{\textsuperscript{89} Judicial College, The Crown Court Compendium: Part I: Jury and Trial Management and Summing Up, June 2023, Ch.10-3 para 18(2)(b)\textsuperscript{90}}

\footnotesize{\textsuperscript{90} Criminal Procedure Rules 2020, 19.2(1)(a)(i)\textsuperscript{91}}

\footnotesize{\textsuperscript{91} T. Bollé et al., ‘The role of evaluation in reaching decisions using automated systems supporting forensic analysis’, (2020) 34 Forensic Science International: Digital Investigation 301016}
or intelligence reliant on computer science CJS focused research, development and operationalisation.

**Approaching AI/ML artefactual evidence with interdisciplinary insight**

AI/ML-applications are typically developed in numerous stages (over two decades for automated facial recognition (AFR)) and at multiple sites. This is all (including false starts and problems in achieving accurate results) recorded in the vast body of general computer science/technological literature. In practice, however, that literature is unlikely to be accessed by many criminal justice professionals.

The technological and scientific papers that record and present such developments are structured differently to legal literature and often contain detailed statistical data to substantiate the results. Such cultural inhibitors to interdisciplinary understanding predate AI/ML. When forensic science practice and statistics began to converge over the reform of fingerprint comparisons, a distinguished statistician, referred to ‘two communities divided by an apparently common language’. 92 Similar cultural inhibitions have been noted more recently in cybercrime studies. Techno-epistemic networks of experts (such as computer and data scientists, both in academia and in cybersecurity companies) have great digital capital in cybercrime research but may lose sight of its ‘socio-technical nature’. 93 Within medicine, early during the Covid-19 pandemic, concerns were expressed about the relationship between quantitative research scientists engaged in COVID-19 clinical trials and the AI/ML community. 94 The practical consequence of the interdisciplinary knowledge gap is that ‘legal personnel have typically struggled to incorporate the advice and insights of mainstream scientific and technical organisations into their consciousness and practice.’ 95 Conversely, the format and sub-disciplinary structure of legal literature must be a barrier to many technologists understanding the jurisdictionally specific legal requirements that their programming must be tailored to achieve. While within digital forensics – an obvious interface for computer science and the law - relevant articles are not necessarily helpfully sign-posted, peer reviewed, or Open Access and tend to deal with ‘isolated forensic challenges ’. 96

Where AI/ML issues are directly addressed in the socio-legal literature, including within ‘the recent burgeoning of American techno-legal studies’, 97 AI/ML reliant predictive policing (deployment, bail and sentencing decisions) 98 has received much greater attention than probative issues. Academic work dealing with expert scientific evidence 99 has focused on the applied sciences and medical

---

92 B. Silverman, then RSS President, in ‘Discussion on the paper by Neumann, Evett and Skerrett’ (2012) 175 Journal of the Royal Statistical Society: Series A (Statistics in Society) 371, 398
95 See Edmond et al., above n.35 764, 788
96 G. Horsman and A. B. Mammen, ‘A glance at digital forensic academic research demographics’ (2020) 60 Science & Justice 399 (Their publication search criteria may have missed much of the relevant legal and computer science research relevant to this article.)
98 See, for example, D. Hunter et al., ‘A Framework for the Efficient and Ethical use of Artificial Intelligence in the Criminal Justice System’ (2020) 47 Fla. St. U.L. Rev 749
99 See, for example, P. Roberts and M. Stockdale (eds), above n.28
spheres,\textsuperscript{100} and like the relevant caselaw is overwhelmingly common law,\textsuperscript{101} and for that matter American. The nuanced manner by which different jurisdictions take note and, up to a point, borrow from each other may not be readily apparent to computer scientists looking for clear and universally standardised rules with which their research must comply.

The importance of interdisciplinary insight can be illustrated in the rest of this section by examples of unreliable or unlawful AI/ML research and operationalisation.

We might like to think that some ‘forms of evidence have unfortunately come and thankfully gone, including, phrenology’.\textsuperscript{102} However, two research studies about predicting criminality from facial appearance appeared in 2017 and 2020. Both reported a high level of accuracy at the proof-of-concept stage. Wu and Zhang recorded cross-validation accuracy of 97\% with a data set of 1,856 facial images.\textsuperscript{103} Hashemi and Hall reported the same score with one of the classifiers used, but against a data set of 44,713 facial images and that their results were ‘not biased to put people of a specific gender or race in a specific category while ignoring their criminal tendency.’\textsuperscript{104} The latter paper was quickly retracted, but solely because the research involving human biometric data had not received institutional ethics clearance.\textsuperscript{105} Presumably as result of this the authors did not respond to criticism of their research.

The research concept and the reported high accuracies were criticised as illusory.\textsuperscript{106} These responses originated in the computer science community, but drew on a combination of interdisciplinary knowledge including pertinent sociological and ethnographic insights (modified and slightly expanded in the summary of some of the issues here):

- \textit{Technical robustness}: the exceptionally high accuracy of the ‘proof of concept’ claims in the two articles could reflect research design errors, such as the programme’s ability to spot differences in metadata (e.g., comparator images may have been standardised as grey scale photographs) rather than any inherent differences between the images themselves.
- \textit{Socio-legal error}: AI/ML tools need to be jurisdiction specific because of variations in the social construction and temporal definition of ‘crime’ and ‘criminal’. For example, how possession and use of marijuana varies under US state laws, and that the decriminalisation of such behaviour is gaining traction. As Wu and Zhang acknowledged, a court conviction was not a reliable method for distinguishing between “criminal” and non-criminal data sets.\textsuperscript{107}

\textsuperscript{100} For example, the National Academy of Sciences (NAS), \textit{Strengthening Forensic Science in the United States: A Path Forward} (National Academies Press: Washington DC, 2009) allocated 59.5 pages to scientific and clinical disciplines compared with 3.5 for digital forensics.
\textsuperscript{101} J. Vuille, ‘Forensic science evidence in non-adversary criminal justice systems’ in P. Roberts and M. Stockdale (eds), above n. 28, 354
\textsuperscript{102} (‘Phrenology’: criminal types or dispositions could be identified by the shapes of their skulls.) See Nutter, above n.32 919, 923
\textsuperscript{103} Xiaolin W. and Xi Z., ‘Automated Inference on Criminality using Face Images’ (2017) \textit{arXiv:1611.04135v3} [cs.CV]
\textsuperscript{104} M. Hashemi and M. Hall, ‘Criminal tendency detection from facial images and the gender bias effect’ (2020) \textit{J Big Data}
\textsuperscript{106} K. W. Bowyer et al., ‘The “Criminality from Face” Illusion’ (2020) \textit{1 IEEE Transactions on Technology and Society} 175
\textsuperscript{107} See Xiaolin W. and Xi Z., above n. 103, Response 1
- **Ethnographic error:** Wu and Zhang argued that the high accuracy of the results was possible because all the images were of individuals of the ‘same race’. 108 This consistent with how differences in the accuracy of different AFR systems reflect skin tone and gender bias in training data sewtetc. 109 They failed to acknowledge the equally high accuracy reported by Hashemi and Hall, who had used highly diverse US data sets. Their response, however, revealed a failure to distinguish between observable variations in facial appearance and, what is now recognised to be a social construct, race. 110 Racial or ethnic categories are socially fluid labels, often based on a less-than-fully transparent combination of self-identification or official ascription111 and, while risk of appearance etc. bias has to be managed to avoid discrimination in many areas of research, is not a source of empirically consistent reliable information for AI/ML data training.

- **Incompatibility of the original concept with a critical area of scientific consensus:** Wu and Zhang’s did not accept that physiologically and anthropometric theories of criminal appearance had long been discredited; 112 the research concept also confused psychological research into social perception of faces with the accuracy of such perceptions. 113

There was also a public law issue in all EU and UK jurisdictions, not necessarily those where the research took place. Wu and Zhang’s ‘non-criminal’ subset consisted of 1126 images acquired without consent from the Internet. 114 Similar activity but on an industrial scale and involving more than 600 law enforcement agencies globally115 searching investigative facial images against images of known individuals harvested without consent in vast numbers from global social media by a commercial AFR developer, Clearview AI Inc. This resulted, inter alia, in data protection proceedings in Canada, Australia and other jurisdictions, with a £7,552,800 fine in the UK. 116 As noted in the

---

108 Ibid., Response 3
109 False positive rates for 189 commercially available AFR tools tested by the US National Institute of Science and Technology (NIST) were found to vary for different demographic groups depending on the geographical location of the supplier: algorithms developed in the US tended to have higher rates for West and East African and East Asian people in one-to-one matching, whereas for many algorithms developed in China this effect was partly reversed, with low false positive rates on East Asian faces. See: NIST, Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects, No 8280 (2019)
112 P. Mazzarello, ‘Cesare Lombroso: an anthropologist between evolution and degeneration’ (2011) 26 Functional Neurology 97
113 See K. W. Bowyer et al, above n.106 at 6.
114 See Xiaolin W. and Xi Z. above n.103 at 3
Canadian and the Australian determinations, 100% accuracy claims were included in the marketing. Law enforcement agencies presumably attracted by such claims seeing the application as highly economically efficient and investigatory effective, either paid for access or tested the application (including in live investigations) in free trials.¹¹⁷

Yet most law enforcement officials did not understand how the technology actually worked. Nor, ... did anyone know much about the company behind the technology.¹¹⁸

This article is focused on criminal proceedings and the need to ensure that CrimPD are used effectively, but this scandal is a reminder of the wider risks within the criminal justice system generally, where potential judicial safeguards are non-existent. The number of individuals wrongly associated with serious offending because of a lack of technological understanding within law enforcement and thereby socially stigmatised, or enrolled in suspect or safeguarding records is unlikely ever to be known.

**Expert witnesses as ‘explainers’ of their AI-dependent findings and participants in AI/ML research and development**

Good medical practice provides helpful guidance about the objectives that expert witnesses should be trained to achieve - consistent with the ultimate issue rule – when explaining the significance of the artefactual or artefactually dependent nature of their evidence in the instant case. Guidance about explaining AI-assisted decisions published by the ICO and Turing Institute in 2022 ¹¹⁹ illustrates how explanations should be given to patients in a high impact (life/death) situation. It is essential that they should understand how the diagnosis was made, including reliance on an AI/ML system. The explanation needs to be intelligible to patients who may not know how to query an AI/ML system output, by discussing for example:

- The quality of data processing: how the data used by the application was used, collected and cleaned and why it was chosen to train the model. Also, information about safeguards to ensure it was accurate, consistent, up to date, balanced and complete.
- What is known about the applications’ performance metrics in terms of the available training data, and the healthcare organisation or third-party vendor that decided how accuracy should be assessed.
- Safeguards to ensure the system’s robustness and reliability if used outside laboratory-controlled conditions.

When providing this information doctors should ‘indicate how much confidence they have in the AI system’s result based on its performance and uncertainty metrics as well as their weighing of other clinical evidence against these measures.’ ¹²⁰

Such an approach does, however, contradict thinking in some influential circles about the transformational nature of AI/ML within criminal justice. For instance, The US President’s Council of

---

¹¹⁷ See Purshouse and Campbell, above, n.115 at 213; In the USA, for example, the Justice and Homeland Security departments, US Drug Enforcement Administration, the FBI, the US Immigration and Customs Enforcement, the US Secret Service and hundreds of police departments, and administrators and campus police at and school scores of universities, P. Dauvergne, *Identified, Tracked, and Profiled* (Edward Elgar: Cheltenham,2022), 60

¹¹⁸ Ibid. at 61 and 63

¹¹⁹ See ICO and Turing Institute, above n. 13

¹²⁰ Ibid. at 12
Advisors on Science and Technology (PCAST) suggested in 2016 that forensic analyses could be performed by an automated system or human examiners exercising little or no judgment. Such a view does not, of course, is unlikely to comply with UK and EU data protection law, that variously provide for controls and remedies against ‘significant decision based solely on automated processing’. 

Although theoretically and doctrinally strong, the adoption of this approach also needs to overcome major practical limitations. Considerable investment is taking place in medical AI/ML applications under the guidance or scrutiny of multi-skilled research and globally interconnected development teams that, inter alia, improves the medical profession’s ability to explain the reliability of AI/ML generated reports. Physicians can view AI/ML generated data critically, for example, by seeing the risk score for a given source of information that contributes to a multiple source prediction to identify potential errors. Expert witnesses, may often use applications that are comparatively rudimentary. Criminal justice is a much smaller and, in terms of economic theory, imperfect market. Until institutional investment in criminal justice AI/ML applications produces sufficiently transparent, detailed and comprehensive information about potential risks. Significant qualifications may have to be given about the reliability of artefactual or artefactually dependent evidence, including, for example, when with hardware changes and the black box issue (both considered below) it may be impossible to judge how reliable the system was in the instant case.

**Evaluation, reliability, accuracy and error**

“Accuracy, ... is partly a question of objective facts and partly a function of striking an appropriate balance for the purposes at hand between tractable generalisations and exhaustive technical detail”

The above comment by an interdisciplinary group of authors (statistician, legal academic and forensic scientists) in a publication about uncertainties statistics and probability, applies equally to computer science. It is germane to explaining the reliability of artefactual or artefactually dependent expert evidence.

AI/ML offers the prospect of standardised and transparent statistical measurements and probability estimates for elements of expert evidence that are at present entirely subjective, especially feature comparisons (i.e., the measurement of latent fingerprint image quality and the probability of it corresponding to other fingerprint data, whether other latent or reference). Without the expected new AI/ML applications, the best that can be achieved for many feature comparison disciplines is to compare variations between different practitioners. Such error rate measurements are helpful in exposing methodological/conceptual flaws and reasons for biased results, but cannot guarantee the avoidance of significant error. Proficiency testing, at least if not undertaken blind and replicating

---

124 See C Aitken et al, above n.85 at 9
casework level difficulties, may have a limited value,\textsuperscript{126} and all the experts tested could have made the same erroneous decision.\textsuperscript{127}

Explanations of the accuracy of such new AI/ML tools and how this is determined by the quality and use of the training data, however, are never short and straightforward. Contrary to the impression created by marketing, numerous metrics are used for evaluating the performance of AI/ML applications, but no single measure is generally superior. Usually, combined metrics are required to gain an understanding of the credibility, validity, reliability, and generalisability of a tool’s performance.

The starting point for such enhanced competency lies in understanding how the classification model or ‘classifier’ sets parameters for AI/ML coding evaluation. In data science, an AI/ML model depends on the performance of the algorithm selected for its development. Common classification tasks, such as image recognition, can use algorithms such as support vector machine (SVM) or convolutional neural networks (CNN) for ML and AI models, respectively. There are dozens of algorithms available for classification and other tasks involving big data. Despite their different composition, they are all designed to deal with factors such as time complexity, scalability, update capability, capacity for generalisation, accuracy, degree of reliability, resilience, and potential impact on validity and verifiability. This is not the place to attempt a comprehensive summary of such issues, which would soon become out of date. The nature of the issues that we have in mind for an explainer’s tool kit can be illustrated\textsuperscript{128}, however, as follows:

- **How the classifier is trained**: This requires a pre-existing dataset containing a correctly labelled set of samples, for instance, external images of firearms. Ideally, a predictive model should classify 100\% of the samples in the dataset with the correct classification label. However, this does not guarantee that all previously unseen samples will be correctly classified. A training dataset is likely to be incomplete since there will always be new samples that have not yet been trained. Nevertheless, based on the generalisability achieved by classifying previous samples, it might be capable of classifying an unseen sample correctly.

- **How the classifier’s accuracy/reliability is demonstrated**: One or more of the different evaluation algorithms that are generally available for computer science research is used to assess how many samples in the existing dataset it was able to classify correctly. By dividing the dataset into a training dataset and a testing dataset, usually in a split of 80\% and 20\%, respectively, where the model has never seen the testing dataset, it should be able to generalise its classification model to classify those test samples. Since 100\% of the dataset is available, the number of correctly/incorrectly classified samples from the testing dataset can be examined, and the different metrics from each evaluator can provide a range of accuracy measures.

- **How the classifier’s accuracy is reported**: If the classifications are all correct, the data science metric “accuracy” is 1.0 (or 100\%). That is simply the fraction of the correctly classified samples out of the total number of classifications. Such measurements, however, are

\textsuperscript{126} See Gardner and Neuman, above, n. 42

\textsuperscript{127} I.E. Dror et al., ‘Biasability and reliability of expert forensic document examiners’, *Forensic Science International* (2021) 318 110610

sensitive to outlying values in unbalanced datasets and for that reason can sometimes be misleading. For example, if the dataset consists of 1 malign value and 99 benign values, the accuracy will be reported as 99/100, (0.99), for example, if the model predicts 100 values as benign by missing the outlying false value which might be a critical one. Accuracy scores can, however, be made more reliable by calculating a balanced accuracy score. The balanced accuracy score calculates the true negative (TN) and true positive (TP) rates, namely, the TP / (TP + false negatives (FN)) and TN / (TN + false positives (FP)), respectively and divides them by two. A balanced accuracy score for the example of 100 benign predictions in a dataset of 99 benign samples and one malign sample would be 0.50. Such reporting is consistent with long-recognised criminal justice expert good practice avoidance of presenting accuracy as a singular number. 129

- **A clearer understanding of accuracy by calculating sensitivity/precision and specificity/recall values:** In data science - as with clinical research and forensic genetics130 - these measures are commonly used and are equally important to accuracy scores, but possibly with discipline-specific differences and terminology (precision and recall in computer science). Precision, or positive predictive value, can be described as the classifier’s ability to correctly label the positive predictions (i.e., the true positives divided by the true positives and false positives). Conversely, recall measures the classifier’s exclusivity in labelling the actual true values (i.e., true positives divided by true positives and false negatives).

Such a toolkit could be of particular use to experts giving evidence in criminal proceedings both as a means of establishing their competence to testify and the reliability of the evidence. In England and Wales, an expert’s report should contain “details of the expert’s qualifications, relevant experience and accreditation”131 as well as “such information as the court may need to decide whether the expert’s opinion is sufficiently reliable to be admissible as evidence”132.

It is essential in the criminal justice context to stress that precision and recall are likely to be essential metrics. They are complementary to accuracy measurements because they are less sensitive to skewed datasets. Precision and recall should be viewed as a pair to give a clear view of the evaluation of the classification model’s performance on the dataset. In cases where both precision and recall values need to be considered, the F-score (sometimes ‘F1-score’) is applicable. The F-score is a harmonised mean of precision and recall that presents a type of average of the two metrics.

Although such evaluation algorithms are able to compensate to some extent for skewed datasets, balanced datasets are preferable. It is possible to create a more reliable dataset and a more reliable model. The dataset can be shuffled like a deck of cards to spread the samples more evenly over the set. It can also be balanced by adding samples of the minority class, for example, by adding 98 malign samples to a dataset consisting of one malign and 99 benign samples. In addition to balance the dataset used for training an ML/AI model, usually, the same fraction of the dataset is not used over and over again as the testing dataset could potentially bias the model given that the training dataset contains only one type of values, and the testing dataset contains only other types of values. In a small dataset of 100 samples, suppose 90 are pictures of hardware tools (hammers, wrenches and such) and ten (10) samples depict firearms. Given that the same ten samples of firearms are used as the testing dataset, the model is trained only to classify hardware tools, resulting in a poor

129 See Nutter, above n.32 at 933
130 See Aitken et al., above n. 85
131 CrimPR, 2020, 19.4(a)
132 CrimPR, 2020, 18.4(h)
accuracy score for identifying firearms. To achieve a more precise performance evaluation score on an imbalanced dataset, the dataset can be divided into “n-fold” partitions to be used in a “cross validation” - meaning the dataset is split into n (usually five or ten) partitions and every partition will serve for example, as the testing partition once; cross-validating the dataset partitions. N-fold cross-validation means a testing dataset will not constantly be the same fraction of the dataset – i.e., not only the same ten samples of firearms as in the previous example - and hence the dataset training the model becomes more nuanced and ultimately results have greater accuracy.

Expert witnesses have been expected to pay particular attention to the relationship between sample size, in which we include training datasets, and potential inaccuracy. 133 ‘The ‘power’ of machine learning in recognising patterns is proportional to the size of the dataset, the smaller the dataset, less powerful and less accurate are the machine learning algorithms.’ 134 Kokol et al have suggested that the solution to this problem, which affects many activities outside criminal justice, might be for learning to be generalised on data sets from various fields so that many different small data sets might become a big data set. While this may be feasible for many types of now automated economic activity, such as contract reviews and financial audits, it may not be a practical or, recalling the risks revealed by the Clearview example, legal way forward for sensitive personal data within the criminal justice context. 135 There is also the issue of legal variation between jurisdictions, to which we shall return in the final subsection.

Technological and technologist anticipative issues

The critical importance of training data has been indicated above. Beyond the proof-of-concept stage, a large and diverse data set must be used for training the programme and predictions need to be tested using data that was not used in any way during model training. There is considerable knowledge about the problems caused by data (often termed ‘algorithmic’) bias during AFR development. White males were over-represented in initial datasets used during the training stage and the images used had been created on film whose chemical formulae was designed to produce sharper images with light skin tones. Unaware of this, programmers did not anticipate how accuracy would be skewed for non-light toned people. 136

Today, computer scientists have a better understanding of the causes of inaccurate or biased performance. Some problems may be missed for years, for example, learning reinforcement. This can occur when the programme is trained to invent ways to accomplish tasks, in effect, by penalising it or rewarding it for achieving specific objectives. It may respond by ‘wireheading’/inventing ‘short cuts’. Background cues or scene biases in the dataset may create shortcut opportunities to recognise the primary objects or may arise from the source or acquisition or preparation method of the data samples. Programmes may learn from the presence or absence of ancillary tokens present in images, including the presence of originator logos or the position of such logos video frames of pornographic images, to classify images on that basis and not the content of the image itself. 137

133 See PCAST, above n. 116
134 P Kokol et al., ‘Machine learning on small size samples: A synthetic knowledge synthesis’ (2022) 105 Science Progress, at 2
135 Ibid, at 10
137 M. Borg et al. ‘Detecting and ranking pornographic content in videos, (2022) 42 Forensic Science International: Digital Investigation 301436
was only recognised as a general problem in 2016 but publications discussing examples of this phenomenon can be traced back to 1983. Other causes of technological risks include:

- **Inadequate scalable oversight**: the programme’s continued adherence to the intended objectives needs to be frequently evaluated during programme training.
- **Robustness** of the programme in operational conditions: ‘harsh real-world conditions’ need to be modelled and tested during the training process. Classification thresholds (the measurable amount of correlation required for a classification to be recorded) set by the programmers may not allow for typical variations in environmental variations (e.g., lighting and camera quality) that affect critical inputs.
- **Pre-operationalisation validation testing parameters**: the reported accuracy of proof-of-concept or ‘developmental validation’ only hold good for objectives set within the parameters for laboratory testing.
- **Programme upgrades**: upgrades that provide more functions or remedy identified defects may result in new modes of operation that have not been previously tested.

While this paper focuses on risks intrinsic to AI/ML programming, however, *hardware changes* may be equally significant and possibly more likely to go unremarked. Changes such as memory size or available disk space may modify the programme’s operation or cause it to behave unpredictably. ‘There is not even a theoretical technical solution to this drawback that will lead to reliable practical countermeasures.’

It is difficult to see how a court can be satisfied about the methodological soundness with which artefactual or artefactually dependent evidence is produced, without the expert witness being able to produce and explain - in the terminology used by the ICO and Turing Institute - evidence-assurance documentation any relevant issues considered immediately above. The problem here is whether defence counsel is aware of the risks described in this article and how they may apply to the instant case. The rarity of reported CrimPD Part 7 challenges, suggests they are not.

*Anticipating how end-users understand and operate AI/ML systems*

Tschider noted from the techno-clinical literature that ‘two of the most crucial choices an AI designer makes are the mechanisms for immediate feedback and correction. If ‘a system trains on data from hospitals with a high degree of resources - such as the newest technologies and the most highly trained practitioners - the model the AI system creates will be oriented towards high-resource use and may not be as effective as one trained on low-resource environments.’ To avoid this

---

138 See the Specification gaming examples in AI - master list : Sheet1 < Specification gaming examples in AI - master list - Google Drive > on 4 June 2022
139 D Amodei et al. ‘Concrete Problems in AI Safety’ (2016) arXiv:1606.06565v2 [cs.AI]
140 See ICO and Turing, above n. 13 at 37
141 NIST, above n. 104 at 6
142 T. Lerer ‘Check the sources: Why secret computer code matters and how defense counsel can get it’, The Champion (21 December) 2021 14
143 See Mason, above n.19 at 130 and 133
144 See Ladkin et al., above n.25 at 2
145 As a minimal digitally-aligned adjustment of CrimPD Pt. 7 (d) and (i) : respectively, ‘if the expert’s opinion relies on the results of the use of any method (for instance, a test, measurement or survey), whether the opinion takes proper account of matters, such as the degree of precision or margin of uncertainty, affecting the accuracy or reliability of those results’; and ‘whether the expert’s methods followed established practice in the field and, if they did not, whether the reason for the divergence has been properly explained.’
146 See ICO and Turing, above n. 13 at 36
disparity, training data should be representative of the population or community where the AI might be used. This suggests, at a minimum, that end user involvement at the inception of an AI/ML project, ideally at the proof-of-concept stage and no later than the development stage is critical for any evidence-assured system. In criminal justice, the kind of experience, behaviours and risks that expert collaboration would enable software developers to anticipate include:

- **Expert knowledge and manipulation of existing systems**: By the early 1980s, at least in parts of the USA, fingerprint examiners specifically tailored their latent print annotations when encoding data in line with observations about variations in the responsiveness of different propriety black box AFIS programmes to metadata variations in input data. Such expertise and the professional culture that gave rise to such behaviour needs to be understood by programmers as early as possible during system development.

- **Expert competence**: increasingly professional and organisational competency in the production of expert evidence is quality assured, but, as noted earlier, for well understood reasons in England and Wales the FSR is having to concentrate on remedial responses to known problems and progress is necessarily slow. Whilst quality assurance of professional and organisational competence will not establish an expert’s competence to give evidence in criminal proceedings per se it may form a part of the basis for making such a determination. A US sentencing case, involving a predictive incarceration issues report, illustrates the extreme end of the risk continuum. The report was skewed by arithmetical error, double counting and conclusions not supported (even after input errors) by the ML-encoded tool outcomes. More deeply seated problems within the institution where evidence is produced or commissioned, however, are more difficult to identify and assess. NIST has begun to trial a practitioner competency assessment methodology to measure this both individually and with reference to demographic characteristics (workplace environment, education, and work experience). Their methodologically, however, has yet to be proved to be successful and initially it only covers mobile and hard-drive forensic investigation. It is certainly unlikely to equal the obligations placed on individuals under Part 7 CrimPD if counsel are sufficiently knowledgeable and resourced to activate these safeguards in relevant cases.

**How would the ‘explainer’ approach resolve the Blackbox/access to source code issue?**

The degree of accuracy and predictability embedded in the operation of source code is critical for the reliability of artefactual or artefactually dependent evidence:

> The code dictates which tasks a computer program performs, how the program performs the tasks, and the order in which the program performs the tasks.  

---

147 C. A. Tschider, ‘Beyond the "Black Box”’, (2021) 98 Denv. L. Rev. 683, 698

148 See NAS, above n.100 at 10-1- 10-7

149 In England and Wales by demonstrating compliance with ISO/IEC 17020/25. Because these were designed for forensic purposes, they need to be supplemented by the FSR’s Code of Practice and Conduct. The latter is an extensive, expanding and frequently updated body of guidance but coverage is incomplete, and progress has to be slow because of the considerable amount of consultation required.


151 See Imwinkelried, above n.64 at 99

152 Imwinkelried, above n.64 at 99
The explainer approach would require candour about the extent to which its accuracy and predictability can be relied upon, and whether with some applications this can be known. Three considerations need to be borne in mind, although - except in the USA - the first can be readily dismissed:

1. The primacy given by American courts to the protection of the commercial confidentiality of source code is explicable as an historical and inconsistent anomaly within all US criminal jurisdictions. The Federal Commerce Clause\textsuperscript{153} advances substantive commercial objectives, at the cost of “some sacrifice of availability of evidence”.\textsuperscript{154} Some US courts have found ways round this restriction, beginning apparently with a Minnesota decision in 2007. Because the procurement process had resulted in a bespoke model that was judged to be public property, trade secrets protections did not apply. The value of this precedent can be seen from a later Minnesota decision (2012). This case revealed how a generally reliable intoximeter malfunctioned when testing the blood of women over sixty and at certain temperatures.\textsuperscript{155} Similarly, access in New York (2021) to the backbox of a probabilistic genotyping tool used to interpret complex DNA evidence was allowed because it had been developed and used by a major public sector agency.\textsuperscript{156} That litigation exposed such significant flaws in this tool that it had to be withdrawn, but by then it had been used ‘in thousands of criminal prosecutions over several years’.\textsuperscript{157}

2. The discretion available to American courts to protect law enforcement/security apps source code to prevent criminal counter measures being developed. In extensive fair trial litigation during 2015 -17 – the majority of requests for defence access to the source code of ‘network investigative techniques’ (NITs), hacking malware, used by the FBI in a child pornography investigation appear to have been refused. Otherwise – in a move familiar to English courts – the prosecution was abandoned.\textsuperscript{158}

3. The impossibility (in practice) and commercial resistance to guaranteeing to a legal standard the correctness of an application’s outputs was explained in the 2020 paper prepared for the Law Commission. Ladkin et al. – all distinguished computer scientists – acknowledged that for any ‘moderately complex software-based computer system’, even book-keeping/transaction-processing systems, it is practically impossible to guarantee the correctness of every software operation. This had to be achieved for safety critical systems, such as aircraft control systems, but the cost of and rarity of scientists able to undertake the mathematical-logical analysis methods (called formal methods) needed to achieve this would be prohibitive for most software development. They attested to the considerable commercial resistance, even in the development of such critical systems, to any requirement for use of such methods.\textsuperscript{159}

4. ‘Dynamic inscrutability’ is a term introduced by Tschider to describe how even a system’s creators may not fully understand how a tool works. In unsupervised machine learning, as the algorithm continues to learn from new data, it is likely to continue to evolve so that a point in time description of the algorithm may be impossible. Complexity is compounded

\textsuperscript{153} U.S. Const. Art.1 s. 8, cl. 3
\textsuperscript{154} See Stein, above n.2 at 93-94 and 101-103. As result of this, for example, Minnesota and New York mandate the courts to recognise the validity of ‘approved’ intoximeters, see Imwinkelried, above n. 63 at 106-108
\textsuperscript{155} See Nutter, above n. 32 at 941
\textsuperscript{156} See Lerer, above n. 142
\textsuperscript{158} See Chen, above n.75
\textsuperscript{159} Ladkin et al. above n.25
when the programme also uses neural networks, or deep learning systems because specific weightings are added to relationships between data elements. Even where an explanation is possible, she suggests that it may not provide ‘the kind of information needed to actually evaluate risks of unfairness, discrimination, safety, or other social impacts.’

This view is shared by many computer scientists. Imwinkelried has suggested a judicially managed two or three stage process for resolving source access disputes. Firstly, the defence must convince the court that the validation information available for the tool (i) ‘do not adequately address the effect of a specified, material variable or condition present in the instant case’ and (ii) could plausibly affect the verdict. Secondly, if that bar is passed, up to two more steps should follow: (a) a new validation study focused on the instant case issues by a defence expert and, if that does not resolve significant expert disagreement, (b) an opportunity for the defence team to examine the source code to assess the accuracy of the results cited in the prosecution case. England and Wales, however, lacks a judicially managed expert evidence dispute resolution procedure. Such a statutory framework was suggested by the Law Commission in 2011 but ultimately rejected on cost grounds. The alternative approach adopted, as discussed above, was the introduction of amendments to the CrimPR and the associated CrimPD, including the introduction of criteria to assist the court with the pre-trial assessment of reliability and making provision for inter alia pre-hearing discussions between experts. Therefore, the jury is left to determine the potential effect of the AI/ML system operation on the evidence, but this will only be possible in practice, if sufficiently comprehensive explanation are either provided by the expert witness(es).

The need for expert witnesses/end-users to participate in AI/ML research and development: a jurisdictional specificity and an illicit trading example

The relationship between jurisdictional specificity and the reliability of AI/ML applications has been noted already in this article. What constitutes unlawful behaviour (a) may change over time (e.g., marijuana decriminalisation), (b) differ significantly between jurisdictions (e.g., the scope to allow marijuana sales in the Netherlands but not the UK) and (c) substantive elements of the offence may vary within jurisdictions (in the Netherlands marijuana can be lawfully sold only to Dutch citizens). Similarly, some cryptomarket vendors ‘offer firearms that have different legal statuses based on the parties’ location and jurisdiction. Bergman and Popov’s PDTOR research has demonstrated how this problem could be resolved by giving end-users access to an annotation tool that would enable them to ensure that automated internet illegal transactions search criteria are jurisdiction specific and easily updated should the law change. tool users would not depend on new releases of the model to maintain the empirical accuracy of data classification. Users could themselves maintain

160 Tschider, above, n.147 at 687-690 and 706
161See, for example, W. Hall and J. Pesenti, Growing the Artificial Intelligence Industry in the UK (BEIS and DCMS: London 2017) 68; and Bryce Goodman and Seth Flaxman, “European Union regulations on algorithmic decision-making and a “right to explanation”” (2016) arXiv:1606.08813v3 [stat.ML]
162 Imwinkelried, above n. 64 at 128-130
163 The Law Commission, Expert evidence in criminal proceedings in England and Wales, (2011), Law Com No.365
165 Now found in CrimPD 2023, Part 7.1
166 CrimPD 2023, Part 7.3
167 J. R. Lee et al., ‘An assessment of the state of firearm sales on the Dark Web’, Journal of Crime and Justice, published online, 06 Apr 2022

24
empirical accuracy, for example, when monitoring illicit firearms trading, adding images of new
types or novel modifications of firearms observed in illicit marketplaces.

This has been published as the results of a proof-of-concept validation research to create an
annotation to improve the reliability of tor cryptomarket surveillance. The operation of this
annotation tool can be summarised as a four-step process:

1. The forensically sound (i.e., chain of custody) record of the manual capture – by an
   investigator - and preservation of a Web page and its metadata as an
   annotated/annotatable data set and the storage of this artefact within an archive.
2. Each artefact is automatically indexed and accessed via a server so that multiple
   investigators can record in the archive their judgement about the artefact’s classification
   using a Web browser that allows full data visualization (i.e., the page as seen on the web
   plus its metadata as originally captured and separately annotations by colleagues).
3. Within the chain of custody record, the archiving of an annotator consensus agreement or
   the statistical calculation of degree of variation between annotator judgements about the
   quality and accuracy of each data set.
4. Relevant annotated artefacts could then be used to create a training data set for the
   unsupervised programming of a web crawler to search the Dark Web and to capture and
   archive ('scraping') additional artefacts that conform to the quality and accuracy parameters
   created and recorded during the annotation process, selected by an AI/ML-based
   classification model.

From a jurisdictional perspective, stage 2 is critical. This would enable criminal justice experts to
ensure that any artefact is only confirmed as evidence of unlawful activity against the substantive
criminal law at time of annotation. These parameters are then embedded for stage 4 when –
theoretically– they cannot be changed, irrespective of how the black box/source code of the crawler
operates.

At this proof-of-concept stage, the results achieved for AI/ML-encoded crawler data capture using
four classification algorithms was a balanced accuracy rate of between 85% and 95% against a small
set of 150 HTML web pages, mostly from dark web marketplaces. The next stage of the research will
involve a bigger data set. It will also test the tool against better protected tor web pages, include
further examination of the resilience of the chain of custody for archived data, and widen
functionality to include the annotation of graphic content located on the Dark Web. The tool could
be incorporated within a criminal justice data system – subject to rigorous connectivity validation -
rather than the free-standing database used for this proof-of-concept research. The tool is highly
adaptable, – as indicated in a later study – it has been proved to be being suitable for use
(including capturing images) with any dark or clear (surface) web crawler and is easily reconfigurable
for Clear (Surface) Web surveillance by using a different browser. The article explains why dark web
scraping – irrespective of the technology used (e.g., multi-threaded distributed crawling engines
used for clear web commercial services) - is significantly slower and labour intensive. This reflects
how ANC network speeds are significantly slower by design and the need for pseudo-random delays
to evade (not always successfully) security features. Slower scraping, however, allows investigators

168 See Bergman and Popov, above n.8; also: O. Popov, J. Bergman, J., & C. Valassi, ‘A framework for a
forensically sound harvesting the dark web’ (2018). Proceedings of the Central European Cybersecurity
Conference 1-7
169 J. Bergman and O.B. Popov, ‘Exploring Dark Web Crawlers: A systematic literature review of dark web
crawlers and their implementation’ (forthcoming) IEEE Access

25
to ‘invigilate’ and steer the process for probative purposes - in a cyberspace location with a deliberately intensified volatility, as servers reportedly disappear regularly from such networks. At the time of writing, what the project lacks is end-user participation, largely because of the pressure of work on criminal justice digital experts.

Conclusions

In this article we set out to explain a significantly different approach to mainstream techno-legal literature for examining the complex and fast changing relationship between law and computer science. An historical inability to adapt to scientific and technologically dependent evidence production is seen primarily as an ethical failure within criminal justice and state institutions. This often arises because of the acceptance of epistemological incomprehension between lawyers and scientists. It is compounded, however, by the political economy of criminal justice and safeguard evasion within state institutions.

In England and Wales doctrine makes a distinction between expert witness competence to give evidence in any circumstances and whether the evidence in the instant case is admissible and, if so, what probative value it carries. The practice of giving expert evidence has been reformed significantly with the CrimPD. Also, since 2023 scientific evidence producers are required increasingly – but not comprehensively - to confirm institutional and specific testimonial conformity with statutory standards that are subject to ongoing revision. This is matched by cultural change within a senior judiciary now committed to supporting the ‘enormous strides in getting forensic science set on a course of absolute science, rather than old wives’ tales or police lore.’

However, such advances in themselves may be insufficient. This caveat is highly relevant to expert opinion evidence that relies on AI/ML applications so that it is either artefactually dependent or wholly artefactual. In such circumstances for expert witnesses to be effectively peritus and to assist the court in determining reliability and decision makers with assessing the weight of evidence it is not enough for them to be competent to give admissible evidence because of their knowledge of a field of forensic activity, for example, forensic genetics or digital forensics. They must also be able to describe potential risks or weaknesses in their evidence that arise because of the interrelationship(s) developed through computer science between their own disciplinary expertise and other sciences (not necessarily just STEM disciplines).

Looking beyond England and Wales and at the wider implications of this article, it is simply unrealistic to expect legal professionals – without the proactive assistance of expert witnesses - to have sufficient scientific expertise to ensure in such circumstances that unreliable evidence is deemed inadmissible, and that weak scientific evidence is presented accurately and fairly – that is with all necessary caveats- to the factfinders. Science today is too diverse, both in its theoretical and applied aspects, for professionals in other fields to necessarily identify problems as they arise in individual cases. The interdisciplinary knowledge gap problem will be amplified as criminal justice decisions increasingly become AI/ML- assisted decisions, where in addition to computer science, relevant evidence may require knowledge of other sciences.

Four key principles emerge from our analysis of the risks that arise from expert opinion evidence production that is either artefactually dependent or wholly artefactual:

1. Interdisciplinary insight is essential with opinion evidence coproduction at the interface of law, computer science, and, variously, other STEM disciplines and social sciences.

170 HL Deb 27 February 2023, Volume 828, Col7GC: Lord Thomas of Cwmgiedd (former LCJ)
2. Lawyers and investigators cannot be relied upon to identify significant risks that may affect the credibility that decision-makers, especially factfinders might accord to opinion evidence that might be highly material to the verdict.

3. The ICO/Turing explainer approach to AI/ML-assisted decision-making is highly relevant for (a) framing professional standards for both the producers and users of such evidence and (b), in a way that is broadly adaptable to the jurisdictionally specific doctrinal and organisational requirements.

4. There is an urgent need to develop law public policy and practice on these matters to overcome institutional and cultural tendencies to safeguard evasion, for example, weaknesses arising from SFR in the UK and the likely primacy of commercial confidentiality over fair trial protections in the USA. The explainer approach appears to be potentially valuable even in the different and difficult circumstances created by the US constitution. The article has also demonstrated how an understanding of medical good practice that has evolved for managing the use of AI/ML applications is an important source of insight for both researching and developing/implementing AI/ML safeguards for criminal justice. It has also linked the criteria for expert witness competency and training for fair trial purposes with the value of such experts engaging in critical and transparent collaboration with computer science researchers and developers throughout the life cycle of such AI/ML applications, including the development and validation of later versions of applications introduced into use.