

Chapter 18: Artificial Intelligence as evidence

Daniel Seng⁺

Abstract

The chapter focuses on the use of artificial intelligence as evidence and specifically on the use of information produced by AI – machine learning and its related technologies – that is used to demonstrate an asserted fact for a legal determination as to the truth or falsity of a proposition. Given AI evidence is a type of electronic evidence, the chapter focuses on AI evidence that involves some element of computer-processing using AI and machine learning – that is used for a determination of facts under judicial inquiry. These include computer-generated transaction summaries and business records, drug analysis reports, genomic typing reports and breath-alcohol tests for identification and evaluation purposes, automatic number plate recognition evidence to show vehicular location, movement and time, automated recordings of nautical traffic, and digital forensics for cybercrimes. The fact that AI is a general-purpose technology whose use is pervasive and that the algorithms associated with machine learning are of general application mean that it is more meaningful to first analyze such evidence on the basis that it is evidence in electronic form, before adding to the analysis the considerations that come from the use of AI and its related technologies such as statistics and algorithms. This chapter provides a review and summary of the use of AI as evidence citing cases and issues from numerous applications.

Introduction

You are a teacher teaching at a middle school in a school district. The school district had declared a district-wide goal of letting go up to 15% of teachers who are considered ineffective. To this end, the district used a new teacher-appraisal system that attempts to measure a teacher's impact on student performance, for better or for worse, based on the students' performance on standardized tests over time. In general, this is done by a statistical model developed by a private company that compares the average test score growth of students taught by the teacher with the statewide average for students in that grade or course. Using the statistical model, the software generates a raw score that is recalibrated into one of five "value-added" effectiveness ratings for any teacher, ranging from "well-above" effectiveness to "well-below" effectiveness.¹

⁺ Founder and Co-Director, Centre for Technology, Robotics, Artificial Intelligence and the Law (TRAIL), Faculty of Law, National University of Singapore and Associate Professor, Faculty of Law, National University of Singapore.

¹ Houston Federation of Teachers, Local 2315 v. Houston Independent School District, 251 F.Supp.3d 1168, 1172 (US D.C., S.D. Texas, Houston Division).

Unfortunately, you are one of the teachers graded as being of “well-below” effectiveness. Upon receipt of the notice that you will be let go, you ask to review the source codes and other information underlying the statistical model and the calculations that are made to generate the effectiveness rating. You are, however, rebuffed, on the basis that these are proprietary trade secrets unavailable to you, your teachers’ union and even the school district that licensed the model and software from the private company.²

Are you able to succeed in challenging your termination notice from the school district?

This is by no means a hypothetical example, but represents the facts of *Houston Federation of Teachers, Local 2315 v. Houston Independent School District*.³ Although the magistrate judge ruled that the teachers’ constitutional due process rights had not been infringed – because there was a rational and legitimate governmental basis for the school district’s use of statistics and the software – he allowed the claim that the teachers’ procedural due process rights – their right to an “opportunity to be heard at a meaningful time and in a meaningful manner” – had been deprived.⁴ As the private company claimed proprietary trade secrets over the statistics and software, this meant that the teachers and their union, and even the school district, had no way to verify or audit the raw scores received.⁵ Nor could the union’s expert, who has been given far greater access to the underlying computer codes than individual teachers. In fact, without the overall statewide statistical model, even teachers who track their own students’ performance test scores would be unable to ascertain their own effectiveness ratings.⁶

The *Houston Federation of Teachers* case is not an isolated one. The raw score – “data” – that is sought to be used to justify the decision of the school district is an example of statistically processed information. In fact, machine learning, the basis for modern day artificial intelligence (AI) tools and products, is a marriage of statistics and knowledge representation.⁷ Today, we are witnessing the pervasive use of AI in all sectors of society, especially with the contemporary popularity of generative AI and large language models.⁸ In the legal domain, this ranges from predictive policing⁹ to facial

² Id., fn. 12.

³ Id.

⁴ Id., at 1175.

⁵ Id., at 1179.

⁶ Id., at 1178. See also fn. 38.

⁷ Flach, Peter. Machine Learning: The Art and Science of Algorithms that Make Sense of Data, xv.

⁸ See e.g. Suzanne McGee, Generative Ai and the Law: AI is here already – with the power to change the legal profession, <https://www.lexisnexis.com/html/lexisnexis-generative-ai-story/>.

⁹ See e.g., Randy Rieland, Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?, SMITHSONIAN MAG. (Mar. 5, 2018), <https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337>.

recognition by law enforcement,¹⁰ from bail assessment and criminal sentencing¹¹ to technology-assisted review for electronic discovery,¹² from due-diligence document reviews¹³ to case assessments,¹⁴ legal drafting of memoranda and pleadings,¹⁵ and even rendering of judgments¹⁶ and damages assessment.¹⁷

Many issues and considerations will arise in relation to the applications of AI to the legal sector. These have been dealt with elsewhere in this book. In this chapter, the focus will be on the use of information produced by AI – machine learning and its related technologies – that is used as evidence to demonstrate an asserted fact for a legal determination as to the truth or falsity of a proposition.¹⁸ Furthermore, AI evidence is a type of electronic evidence – “data (comprising the output of analogue devices or data in digital form) that is generated, processed, stored or communicated by any digital device, computer or computer system or conveyed over a digital transmission system, that has the potential to make the factual account of either party more probable or less probable than it would be without the evidence.”¹⁹ Thus, for this chapter, we will not discuss data produced by AI that is used for management and

¹⁰ See e.g., Jennifer Valentino-DeVries, How the Police Use Facial Recognition, and Where It Falls Short, 12 Jan. 2020, <https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html>.

¹¹ See e.g., *Wisconsin v. Loomis*, 881 N.W.2d 749, 755 (2016).

¹² See e.g., David Benson et al, What is Technology Assisted Review and how should it be used in modern discovery? 24 Jul 2023, <https://www.claytonutz.com/insights/2023/july/what-is-technology-assisted-review-and-how-should-it-be-used-in-modern-discovery>.

¹³ See e.g., Chris O’Leary and Raees Nakuhuda, How AI for M&A due diligence is changing every aspect of the deal process, Thomson Reuters, 16 Apr. 2023, <https://legal.thomsonreuters.com/en/insights/articles/how-ai-and-document-intelligence-are-changing-the-legal-tech-game>.

¹⁴ See e.g., Lex Machina, “Using Legal Analytics for Early Case Assessment” (April 2018) https://lexmachina.com/wp-content/uploads/2018/04/Early-Case-Assessment-UseCase.pdf?utm_source=website&utm_medium=datasheets&utm_campaign=resources.

¹⁵ See e.g., *Park v Kim*, 91 F.4th 610 (U.S. CA, 2nd Cir., 2024)

¹⁶ See e.g., People’s Republic of China Supreme People’s Court, Chinese Courts and Internet Judiciary White Paper (中国法院的互联网司法), 5 Dec 2019, <https://www.chinajusticeobserver.com/law/x/chinese-courts-and-internet-judiciary>; Stephen Chen, China’s court AI reaches every corner of justice system, advising judges and streamlining punishment, 13 Jul 2022, <https://www.scmp.com/news/china/science/article/3185140/chinas-court-ai-reaches-every-corner-justice-system-advising>;

¹⁷ See e.g., Singapore Academy of Law, Motor Accident Claims Online Outcome Simulators, <https://motoraccidents.lawnet.sg/>.

¹⁸ “Any knowable fact or group of facts, considered with a view to its being offered before a legal tribunal for the purpose of producing a conviction, positive or negative, on the part of the tribunal, as to the truth of a proposition, on which the determination of the tribunal is to be asked.” John Henry Wigmore, §1, A treatise on the system of evidence in trials at common law: including the statutes and judicial decisions of all jurisdictions of the United States, at 3, Vol 1 (1904).

¹⁹ Steven Murdoch, Daniel Seng and Burkhard Schafer, The Sources and Characteristics of Electronic Evidence in Stephen Mason and Daniel Seng, *Electronic Evidence* (5th ed, 2021) 41, Chapter 1 [hereinafter *Electronic Evidence*].

administration of investigation, legal processes and litigation. We will also not discuss the discovery of electronically stored information via predictive coding.²⁰ Instead, we will be focused on AI evidence – evidence that involves some element of computer-processing using AI and machine learning – that is used for a determination of facts under judicial inquiry.²¹ These would include computer-generated transaction summaries and business records,²² drug analysis reports,²³ genomic typing reports²⁴ and breath-alcohol tests²⁵ for identification and evaluation purposes, automatic number plate recognition evidence to show vehicular location, movement and time,²⁶ automated recordings of nautical traffic,²⁷ and digital forensics for cybercrimes.²⁸ The fact that AI is a general-purpose technology²⁹ whose use is pervasive and that the algorithms associated with machine learning are of general application mean that it is more meaningful to first analyze such evidence on the basis that it is evidence in electronic form,³⁰ before adding to the analysis the considerations that come from the use of AI and its related technologies such as statistics and algorithms.³¹

Characteristics of Electronic Evidence

So, what are some of the defining characteristics of electronic evidence that make their use and management as evidence in a court of law challenging? In *Electronic Evidence*,

²⁰ Also known as technology-assisted review (“TAR”), it uses AI-driven tools to search for and discover the electronic materials required for a party’s case. TAR has received judicial endorsement in the U.S. (e.g. *Da Silva Moore v. Publicis Groupe* 287 FRD 183 (SDNY, 2012)), England and Wales (e.g. *Pyrrho Investments Ltd v MWB Property Ltd* [2016] EWHC 256 (Ch)) and Australia (e.g. *McConnell Dowell Constructors (Aust) Pty Ltd v Santam Ltd (No 1)* [2016] VSC 734). See “The Sedona Principles, Third Edition, Best Practices, Recommendations & Principles for Addressing Electronic Document Production” (2018) 19 Sedona Conf. J. 1, 165.

²¹ See e.g., Singapore Evidence Act, s 3 (definitions of “evidence” and “fact”)

²² *Aw Kew Lim & Ors v Public Prosecutor* [1987] 2 MLJ 601 (Singapore High Court); *Crawford-Brunt v Kruskall*, 489 F.Spp.3d 1 (U.S.D.C., D. Mass., 2020).

²³ *R v Wood* (1983) 76 Cr App R 23; *Public Prosecutor v Ang Soon Huat* [1990] 2 SLR(R) 246.

²⁴ *People v. Wakefield*, 38 N.Y.3d 367 (C.A. of N.Y.).

²⁵ *R v McKeown*; *R v Jones* [1997] 1 WLR 295; *Director of Public Prosecutions v Barber* (1998) 163 JP 457. For a recent example of a U.S. case on breath alcohol tests, see *Commonwealth v. Hallinan* 491 Mass. 730 (Mass. S. Jud. C., 2023).

²⁶ See, e.g., *Jackson v R* [2011] EWCA Crim 1870; Attorney General’s Reference No 114–115 of 2009 [2010] EWCA Crim 1459, Annexure A; *Najib v R* [2013] EWCA Crim 86 at [9]; *Khan v R* [2013] EWCA Crim 2230; *Welsh v R* [2014] EWCA Crim 1027 at [2]; and *Re A (death of a baby)* [2011] EWHC 2754 (Fam).

²⁷ *Sapporo Maru v Statue of Liberty*; *The Statue of Liberty* [1968] 1 WLR 739; [1968] 2 All ER 195.

²⁸ Dipo Dunsin et. al., A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response, (2024) 48 Forensic Science International: Digital Investigation 301675, <https://doi.org/10.1016/j.fsidi.2023.301675>.

²⁹ Ford, Martin. Architects of Intelligence: The truth about AI from the people building it, 2.

³⁰ Daniel Seng and Stephen Mason, Artificial Intelligence and Evidence, (2021) 33 SAclJ 241, 242-243 (hereinafter *AI and Evidence*).

³¹ *Id.*, at 245-247.

the authors noted that a standalone piece of electronic evidence as an “electronic document” does not actually exist. Instead, it is “better understood as a process by which otherwise unintelligible pieces of data that are distributed over the storage medium are assembled, processed and rendered legible for a human user.”³² The electronic document does not exist independently from the software process that creates it from the hardware device.³³

To render the data into human-readable form requires the mediation of a set of technologies. These technologies need to be standardized in order for the same version of the data to be recreated.³⁴ For instance, when Apple controversially changed the “hand gun” emoji into a “water pistol” emoji to reflect its stance on the US gun violence debate, it set off an entire debate about whether these emojis accurately reflected their users’ communications.³⁵

And without further augmentation such as encryption keys and the use of checksums such as hash values, electronic documents are easy to manipulate: they can be copied, altered, updated or deleted.³⁶ This includes deliberate damage or deletion as well as accidental damage or deletion arising from deterioration of the storage media, or corruption of the media or communications from networked storage.³⁷ To compound the issue, electronic documents stored on the network tend to have hierarchical layers of access controls, which means that multiple parties can access the same electronic document, if the access privileges are compromised.³⁸ For instance, a recent Google database leak ironically confirmed that Google inadvertently stored about 1,000 hours of audio recordings of various children as part of its speech service (which were not supposed to be collected without parental consent under the U.S. Children’s Online Privacy Protection Act or COPPA)³⁹, retained video view histories of users who had opted to delete such histories *and* continued to use these records to recommend new

³² *Electronic Evidence*, 43, ¶1.117.

³³ Id. In fact, the failure to appreciate that an “electronic document” is just an item of convenience and instead ascribe to it legal characteristics such as property interests will create additional and intractable legal issues. See Seng, Daniel Kiat Boon and Low, Kelvin F.K., *Data Objects: New Things or No-Thing More Than Ignis Fatuus* (December 21, 2022). (2025) 17 *Law, Innovation and Technology* (forthcoming), Available at SSRN: <https://ssrn.com/abstract=4308631> or <http://dx.doi.org/10.2139/ssrn.4308631> [hereinafter *Data Objects*].

³⁴ *Electronic Evidence*, 44, ¶ 1.118 to 1.120.

³⁵ See e.g., Heather Kelly, *Apple replaces the pistol emoji with a water gun*, 2 Aug. 2016, <https://money.cnn.com/2016/08/01/technology/apple-pistol-emoji/>.

³⁶ *Electronic Evidence*, 46, ¶ 1.125.

³⁷ Id, 48, ¶ 1.129-1.130.

³⁸ Daniel Seng and Kelvin Low F.K., “Data Objects: New Things or No-Thing More than Ignis Fatuus”, (2025) 17 *Law, Innovation and Technology* 1.

³⁹ U.S. Federal Trade Commission, *Complying with COPPA: Frequently Asked Questions*, July 2020, <https://www.ftc.gov/business-guidance/resources/complying-coppa-frequently-asked-questions>.

videos to the users.⁴⁰ In another recent incident, millions of Facebook, Google and TikTok users had their accounts compromised when an SMS routing company leaked its own database that contained supposedly private two-factor authentication codes for these users into public view.⁴¹ After all, the basic premise when a data object is stored for access on a network is resource sharing.⁴²

AI Evidence

Information generated via AI systems is electronic in nature, with some additional characteristics. While non-AI systems work through the provision of manual instructions and rules to deal with different situations, AI systems operate through algorithms and systems that improve their knowledge or performance with “experience”.⁴³ This experience comes from data which must be supplied to the AI system – it must be “trained” on the data and the ML algorithm will automatically detect patterns in the data and generalize from the patterns to “choose” among multiple solutions that are possible.⁴⁴ In other words, the development of an AI system centers around building an “accurate” statistical model of the training data, whereupon it can “predict” or render decisions - generate “inferences” - based on the model.⁴⁵

The Double-edged Situation of Training Data

This yields the first key characteristic of AI evidence: the statistical model is intended to be a generalization of its training data. In other words, the quality of the AI evidence is determined by the quality of its training data and the choices made by the selected ML algorithms. If there are any biases inherent or hidden in the training data, the AI system will learn such biases through these algorithms.⁴⁶ For example, an AI system for determining whether a prisoner should be released by the parole board will exhibit

⁴⁰ Scott Ikeda, Google Database Leak Documents 6 Years of Privacy Breaches, 7 June 2024, <https://www.cpomagazine.com/data-privacy/google-database-leak-documents-6-years-of-privacy-breaches/>.

⁴¹ James Laird, Millions of Facebook and Google Users Hit by 2FA Data Leak, 1 Mar. 2024, <https://tech.co/news/millions-facebook-google-2fa-data-leak>.

⁴² A. Silberschatz, P.B. Galvin and G. Gagne, Operating System Concepts, 10th ed. (Hoboken 2018), 735.

⁴³ Flach, Peter. Machine Learning: The Art and Science of Algorithms that Make Sense of Data, 3; Raschka, Sebastian and Mirjalili, Vahid, Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2 (3rd ed) 39.

⁴⁴ Kevin P. Murphy. Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series) (2013) 2.

⁴⁵ Flach, Peter. Machine Learning: The Art and Science of Algorithms that Make Sense of Data, 3.

⁴⁶ See David Lehr and Paul Ohm, ‘Playing with the Data: What Legal Scholars Should Learn About Machine Learning’, 51 U.C. Davis L. Rev., 653, (2017), 655. See also Solon Barocas and Andrew D. Selbst, ‘Big Data’s Disparate Impact’, 104 Calif. L. Rev. 671 (2016), which is highly recommended by Lehr and Ohm, at 665 and Deven R. Desai and Joshua S. Kroll, ‘Trust But Verify: A Guide to Algorithms and the Law’, 31 Harv.J.L.Tech. 1, 23-35.

racial bias if it has been trained on data that contains such bias.⁴⁷ Another example of such biases would be Amazon's failed experiment with a hiring algorithm that showed bias against women because it was trained on the resumes of its existing pool of male programmers and developers.⁴⁸ As a recent report from the UN opined, tersely, "The gender digital divide creates a data gap that is reflected in the gender bias in AI. Who creates AI and what biases are built into AI data (or not), can perpetuate, widen, or reduce gender equality gaps."⁴⁹

Edge and Corner Cases

While it is undeniable that AI has made tremendous progress in its accuracy and thus concomitantly broadened its application and reach, it is a truism that in certain situations, it can also give results that are unexpected or are incorrect.⁵⁰ These situations are referred to as edge or boundary cases – where the AI system has to operate at an extreme operating parameter, and corner cases – where a situation arises outside normal operating parameters, and multiple parameters or conditions are simultaneously at extreme levels.⁵¹ Whereas in the past, good software engineering practices will respond to such challenges by brainstorming and explicitly coding algorithms that comprehend and implement correct responses, with machine learning and deep learning, the coding process is very different: the AI system deep learning model has to be trained with an extensive data set that encompasses normal operating situations *and* rarely encountered situations.⁵² This ostensibly simple exercise belies serious issues. For instance, image data can be highly varied, as different camera

⁴⁷ *State v Loomis*, 881 NW 2d 749 (Wis. 2016), cert. denied, 137 S Ct 2290 (2017); Susan Nevelow Mart, "The Algorithm as a Human Artifact: Implications for Legal [Re]Search" (2017) 109 Law Libr J 387; Anupam Chander, "The Racist Algorithm?" (2017) 115 Mich L Rev 1023; Molly Griffard, "A Bias-Free Predictive Policing Tool?: An Evaluation of the NYPD's Patternizr"(2019) 47 Fordham Urb LJ 43; Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases" (2017) 356(6334)*Science* 183; Jieyu Zhao *et al*, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", (Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing)(Association for Computational Linguistics, 2017) <https://www.aclweb.org/anthology/D17-1323.pdf>.

⁴⁸ Jeffrey Dastin, Insight - Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 11 Oct 2018, <https://www.reuters.com/article/idUSKCN1MK0AG/>. See also Genevieve Smith & Ishita Rustagi, When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity, Stanford Social Innovation Review, 31 Mar. 2021, https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity.

⁴⁹ Artificial Intelligence and Gender Equality, UN Women, 22 May 2024, <https://www.unwomen.org/en/news-stories/explainer/2024/05/artificial-intelligence-and-gender-equality>.

⁵⁰ See e.g., Will Knight, "AI's Language Problem", *MIT Technology Review* (9 August 2016) <https://www.technologyreview.com/s/602094/ais-language-problem/>.

⁵¹ Wikipedia, Edge case, https://en.wikipedia.org/wiki/Edge_case#Software_engineering; Wikipedia, Corner case, https://en.wikipedia.org/wiki/Corner_case.

⁵² Brian Dipert, Edge and corner cases: AI's conceptual simplifications and potential complications, EDN, 21 Aug 2023, <https://www.edn.com/edge-and-corner-cases-ais-conceptual-simplifications-and-potential-complications/> [hereinafter *Edge and corner cases*].

lenses and settings, image sensors and focal depths can be used. And images of human subjects may have different orientations, be at different distances, have different skin tones, and wear different outfits.⁵³ In such a case, feeding the AI model with an excess of situations such as images and sensors will invariably lead to an unnecessarily bloated model that could slow subsequent operations. In addition, many of these edge and corner situations involve different presentations of the data e.g. a cyclist who is pushing her bicycle instead of riding it,⁵⁴ or children dressed up as inanimate objects during Halloween.⁵⁵ Even then, errors still crop up, such as the recent recall of 600 Waymo self-driving vehicles for striking telephone poles,⁵⁶ or NHTSA's formal probe into Tesla autopilot vehicles crashing into emergency vehicles.⁵⁷ As the Forbes report states:

It is important to understand that these systems, begin driver assist, rather than self driving, are not designed to recognize and see everything. *It is expected and part of the design parameters that they will miss things.* While there are generally efforts to make them better and better at identifying any obstacle on the road, the responsibility for supervision and action still lives with the driver. The driver, however, often starts to see themselves as secondary, or not even needed at all. While one could say, "They simply should not hit things," that is not an option on the table, or is a way of saying, "there must not be driver assist tools."⁵⁸ (emphasis added)

⁵³ *Id.* See Jennifer Valentino-DeVries, How the Police Use Facial Recognition, and Where It Falls Short, NYT, 12 Jan. 2020, <https://www.nytimes.com/2020/01/12/technology/facial-recognitionpolice.html>. For these reasons, among others, some cities such as San Francisco, Somerville and Oakland have banned local law enforcement from using facial recognition technologies. Shirin Ghaffary and Rani Molla, Here's where the US government is using facial recognition technology to surveil Americans, Vox, 10 Dec. 2019, <https://www.vox.com/recode/2019/7/18/20698307/facial-recognition-technology-us-government-fight-for-the-future>.

⁵⁴ NTSB, Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018: Accident Report, NTSB/HAR-19/03, 15-17, <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf> ("ADS [Automated Driving System] changed the pedestrian's classification several times, alternating between vehicle, bicycle, and other. Because the system never classified the pedestrian as such, and the system's design excluded tracking history for nonpersisting objects—those with changed classifications—it was unable to correctly predict the pedestrian's path").

⁵⁵ Jordan Perch, Google's Driverless Cars Are Learning How to Recognize Kids, 10 Mar 2018, <https://greenerideal.com/news/vehicles/1118-googles-driverless-cars-are-learning-how-to-recognize-kids/>.

⁵⁶ Associated Press, Waymo recalls more than 600 self-driving vehicles for software, map updates after one strikes a pole, 13 June 2024, <https://wtop.com/news/2024/06/waymo-recalls-more-than-600-self-driving-vehicles-for-software-map-updates-after-one-strikes-a-pole/>.

⁵⁷ Brad Templeton, Teslas Are Crashing Into Emergency Vehicles Too Much, So NHTSA Asks Other Car Companies About It, Forbes, 20 Sep 2021, <https://www.forbes.com/sites/bradtempleton/2021/09/20/teslas-are-crashing-into-emergency-vehicles-too-much-so-nhtsa-asks-other-car-companies-about-it/>.

⁵⁸ *Id.*

Absence of Fallbacks

The fact that AI systems can and will fail implies that a system of risk management and internal controls to address the risks involved is crucial. When human expertise fails, humans fall back on “common sense”. But AI systems per se do not define or encode this behavior.⁵⁹ “The field of software engineering is aimed at producing reliable software, but the emphasis has historically been on correctness, not safety”.⁶⁰ As such, when an AI system fails, it will fail in a serious or spectacular manner – a condition known as “brittleness”. Examples of these abound – including fastening stickers on stop signs to make an AI misread it,⁶¹ misidentifying the white side of the tractor trailer as part of the brightly lit sky and thus not applying the brakes for the autonomously-driven vehicle,⁶² or failing to do simple math!⁶³

While research has been done to make AI systems behave more human like or more ethically, including mirroring human behavior through observation (also known as imitation learning), this may suffer from a value alignment problem – where the machine will learn from and repeat human mistakes.⁶⁴ Furthermore, in many complex settings that involve social interactions with humans, there is unlikely to be a correct and converged solution to the setting.⁶⁵ As the authors Russell and Norvig quipped, “After all, many humans never quite learn what makes other humans tick, despite a lifetime of experience, and many of us are unsure of our own preferences too.”⁶⁶ Thus this is unlikely to be a problem that can be resolved soon.

Interpretability Issues

Related to all these issues is the challenge of interpretability – the representation of the knowledge encoded in the statistical model in AI systems in a form that is

⁵⁹ Fallback mechanisms are specifically required for autonomous vehicles. See e.g., California Title 13 Division 1, Chapter 1, Art 3.7: Testing of Autonomous Vehicles, § 288.06(c)(2).

⁶⁰ Russell, Stuart and Norvig, Peter. *Artificial Intelligence: A Modern Approach* (Pearson Series in Artificial Intelligence) at 1002.

⁶¹ See e.g., Eykholt et al, “Robust Physical-World Attacks on Deep Learning Visual Classification,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1625-1634, doi: 10.1109/CVPR.2018.00175, <https://ieeexplore.ieee.org/document/8578273>. The science around this is called adversarial attacks.

⁶² NTSB, Collision Between a Car Operating With Automated vehicle Control systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016, Accident Report NTSB/HAR-17/02 PB2017-102600, 15-16, 30-31, <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1702.pdf>

⁶³ Hendrycks et al, Measuring Mathematical Problem Solving With the MATH Dataset, arXiv (2021), <https://arxiv.org/abs/2103.03874>.

⁶⁴ Id., at 1003.

⁶⁵ “Humans sometimes lie or cheat, or do things they know are wrong. They sometimes take self-destructive actions like overeating or abusing drugs. AI systems need not learn to adopt these problematic tendencies, but they must understand that they exist when interpreting human behavior to get at the underlying human preferences.” Id., at 1004.

⁶⁶ Id., at 1004.

comprehensible by humans. Computer scientists do not always know which feature or which dataset influenced the AI decision or prediction, one way or the other.⁶⁷ While progress has been made by computer scientists in the field of explainable AI (or XAI), especially with regards deep learning on neural networks,⁶⁸ achieving adequate understanding is not easy.

All these special characteristics of AI translate into substantial issues regarding the admissibility of AI as evidence that range from excluding such evidence as hearsay to challenges about the authenticity of AI evidence. We now turn to an examination of these issues.

Hearsay

The rule against hearsay provides that a statement that an out-of-court statement made by a declarant that is offered by a party in evidence to prove the truth of the matter asserted in the statement, is generally inadmissible.⁶⁹ It is a rule that is deeply embedded in Anglo-American jurisprudence. Although this rule does not exist in Continental Europe's proof-taking procedures, the rules that emerge from judicial practice in the Continental courts is that "hearsay" evidence is normally not entitled to any probative weight, unless there is some corroborative evidence.⁷⁰ As such, the discussion that follows, although based on Anglo-American jurisprudence, that considers the probative value of AI evidence in the context of the common law hearsay rule, remains of relevance.

The scope of the rule against hearsay may vary from jurisdiction to jurisdiction: some would treat implied or unintended assertions as hearsay, but others would exclude such assertions from the rule.⁷¹ Other jurisdictions have effected statutory reforms that have either limited the scope of the rule or replaced the rule with statutory

⁶⁷ See e.g., Ford, Martin, Interview with James Manyika: Architects of Intelligence: The truth about AI from the people building it, 280-281.

⁶⁸ These are through techniques such as Local-Interpretable-Model Agnostic Explanations or LIME and Generalized Additive Models or GAMs. See e.g., Christoph Molnar, "Interpretable Models: GLM, GAM and more" (Chapter 5), <https://christophm.github.io/interpretable-ml-book/extend-lm.html> and "Local Surrogate (LIME)" (Chapter 9), <https://christophm.github.io/interpretable-ml-book/lime.html> in Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2024).

⁶⁹ See U.S. Federal Rules of Evidence, § 801(c); Crawford v Washington 124 S.Ct. 1354 (2004), 541 U.S. 36 (2004). The common law rule is as stated in the leading case of Subramaniam v Public Prosecutor [1956] 1 WLR 965, 969 (Privy Council on appeal from Malaysia) ("Evidence of a statement made to a witness by a person who is not himself called as a witness may or may not be hearsay. It is hearsay and inadmissible when the object of the evidence is to establish the truth of what is contained in the statement. It is not hearsay and is admissible when it is proposed to establish by the evidence, not the truth of the statement, but the fact that it was made.").

⁷⁰ See Paul B Rava, Comparative Study of Hearsay Evidence Abroad, (1970) 4 International Lawyer 156, 158. See also Doris Jonas Freed, (1970) 4 International Lawyer 159.

⁷¹ See, e.g., s 115(3) of the UK Criminal Justice Act 2003(c 44); s 4 of the New Zealand Evidence Act 2006; and s 59(1) of the Australian Commonwealth Evidence Act 1995.

mechanisms to exclude hearsay statements based on prior notice⁷² or in the interests of justice.⁷³ Another way to bypass the hearsay rule is where evidence is offered on the theory that it is not a statement because it is real evidence. On this theory, AI evidence is machine-generated output rather than intentionally assertive statements made by human declarants. There are no express or implied factual assertions offered for their substantive truth, and therefore AI evidence would not be regarded as hearsay.⁷⁴

Real Evidence vs. Hearsay Evidence

In this regard, it is worth noting the origin of the “real evidence” rule. In a leading English case on this rule, the court in *Castle v Cross* accepted the definition of “real evidence” as “evidence afforded by the production of physical objects for inspection or other examination by the court”.⁷⁵ Analogizing with a photograph of radar reception in *The Statue of Liberty*, the court accepted the submission of counsel that:

the fact that it is a sophisticated machine which depends for part of its operation on what has been described as computer control does not set it in a different class from other sophisticated mechanical devices and instruments of the kind referred to by Simon P in *The Statue of Liberty*.⁷⁶

And in doing so, the court held that the print-out from the breath-alcohol device that read “one no sample” was admissible as real evidence and rejected the accused’s appeal on the charge that he had failed to deliver sufficient breath into the machine to carry out the required analysis.⁷⁷ Courts have likewise accepted as real evidence X-ray spectrometer, neutron transmission chemical analysis and HPLC printouts generated by a computer,⁷⁸ a video recording showing an exchange of drugs between two people,⁷⁹ a satellite image with accompanying GPS coordinates.⁸⁰ In contrast, courts have rejected evidence of photographs from a social networking site and a YouTube video that portrayed the defendant as a member of an organized and violent criminal gang,⁸¹ car manufacturers’ records of the identification numbers of the engine, chassis and cylinder block numbers recorded by the manufacturers’ employees,⁸² and

⁷² See, e.g., s 2 of the England and Wales Civil Evidence Act 1995 (c 38); s 22 of the New Zealand Evidence Act 2006; and s 67 of the Australian Commonwealth Evidence Act 1995 (Cth), s 67.

⁷³ See, e.g., s 114(1)(d) of the UK Criminal Justice Act 2003 (c 44).

⁷⁴ Paul W. Grimm, Maura R. Grossman, and Gordon V. Cormack, Artificial Intelligence as Evidence, 19 Nw. J. Tech. & Intell. Prop. 9, 85-86 (2021).

⁷⁵ *Castle v Cross*, [1985] 1 All ER 87, 89 (Q.B.D.).

⁷⁶ *Id.*, at 89.

⁷⁷ *Id.*, at 90.

⁷⁸ *Wood* (1982) Cr App R 23; *PP v Ang Soon Huat* [1990] 2 SLR(R) 246, [1991] 1 MLJ 1.

⁷⁹ *U.S. v. Wallace*, 753 F.3d 671, 675 (7th Cir. 2014).

⁸⁰ *U.S. v. Lizarraga-Tirado*, 789 F.3d 1107, 1109-10 (9th Cir. 2015).

⁸¹ *Bucknor v R*, [2010] EWCA Crim 1152.

⁸² *Myers v. DPP*, [1965] AC 1001 (House of Lords).

company business registration records,⁸³ holding the photographs, video and records to be hearsay.

The Three Categories of Electronic Evidence

Several points can be made about the legal position in this regard. First, it is not a given that all electronic evidence, and by parity of reasoning, all AI evidence, is incontrovertibly real evidence. Computers are general data processing devices and can be configured to undertake a variety of tasks. They can be used to accept and store human-supplied input and provide for the retrieval of such input. In this form, the electronic evidence that is tendered principally for the testimonial content of the human input. This information may be digitized, compressed, transcribed, indexed and processed in some way before being stored on the device or some other platform. But the processing therein is generally incidental to the use of such devices as recordation devices for the evidential value of human input. Evidence used in this way is generally testimonial. This would explain the treatment of the aforesaid manufacturing records and company business registration records as hearsay. As Lord Reid opined:

[T]he reason why this evidence is maintained to have been inadmissible is that its cogency depends on hearsay. The witness could only say that a record made by someone else showed that, if the record was correctly made, a car had left the works bearing three particular numbers [the engine, chassis and cylinder block numbers]. He could not prove that the record was correct or that the numbers which it contained were in fact the numbers on the car when it was made.⁸⁴

In previous writings by the author, these are described as Category 1 electronic evidence.⁸⁵

Where computers are indeed used to process information, the evidence elicited will be substantially the product of automation. Tape recorders,⁸⁶ digital cameras and video camcorders are examples of devices that generate such evidence. Fundamentally, after some form of manual setup, these devices work autonomously by collecting sound (and light) information via sensors and converting that via processing into digital (or analogue) form. After some additional processing, depending on the specifications and features of these recorders and cameras, the sounds, images or video streams (and their sounds) are stored or recorded. When the sounds, images or videos are retrieved, they are used to replicate the input or environment within which these devices are situated.⁸⁷

⁸³ *Aw Kew Lim & Ors v PP* [1987] SLR(R) 443, [1987] 2 MLJ 601.

⁸⁴ *Id.*, at 1019.

⁸⁵ *AI and Evidence*, at 259.

⁸⁶ *R v Maqsd Ali*, *R v Ashiq Hussain*, [1965] 2 All ER at 469; [1966] 1 QB at 701.

⁸⁷ Daniel Seng, *Electronic Evidence*, in Chapter 23, *Law and Technology* (2022), at §23.009.

Such evidence would generally not be used testimonially. Instead, they would be considered real evidence and not hearsay – or Category 2 electronic evidence in previous writings by the author. The chemical analysis reports, in situ video recordings showing a drug deal between parties and satellite images with GPS coordinates would all be real evidence.⁸⁸

A large majority of electronic evidence however will be a hybrid of Category 1 and Category 2 evidence – which this author has termed Category 3 evidence. In this category, the electronic device functions through a mix of human-supplied input and produces data output without further human intervention. The human input could be in the form of initial calibration of such devices or their internal programming, or some measure of selection, coordination and arrangement. These are ex facie testimonial input of some form in the absence of further elucidation.⁸⁹ Consider AI systems and their reliance on training data: the machine learning algorithms will have variously selected and generalized from the multitude of datasets that are largely in the form of human-supplied input. For the latest multimodal large language models, these range from code samples to drawings, from conversations to writings, from sound snippets to video recordings. In such a case, the line between whether or not the evidence is or is not hearsay can be a difficult one to draw: the difference really is one of degree that represents the relative significance of the level of contribution of human-supplied input and pre-programmed autonomous processes to the eventual output, and the extent to which the testimonial input qualifies the (purportedly) non-testimonial use of the output.⁹⁰ A detailed and careful case-by-case analysis in this regard is required.

To illustrate this, take the case of the photographs and videos from social media sites. Photographs and video recordings are generally considered non-testimonial evidence because they passively record activities that take place in the environment where they are placed. However, in the case social media content, the manner in which the recordings are made, edited and presented would especially reflect the editorial input of the unidentified and out-of-court video editor as the declarant. In such a case where the testimonial (human) input dominates, the English Court of Appeal in *Bucknor* has made the right decision to treat the evidence as hearsay.

Where AI evidence is tendered, it is submitted that this analysis will take on even more significance. Some parts of AI evidence will be hearsay, and other parts will be real evidence. Careful scrutiny of the parts of the evidence is required to discern the difference. We will take this screen capture from Google Scholar about this author and his writings as an example, taking advantage of the fact that this author has personal knowledge of his own academic output:

⁸⁸ *Id.*

⁸⁹ *Id.*, at §23.011. See also *AI and Evidence*, at 261.

⁹⁰ *AI and Evidence*, at 261.

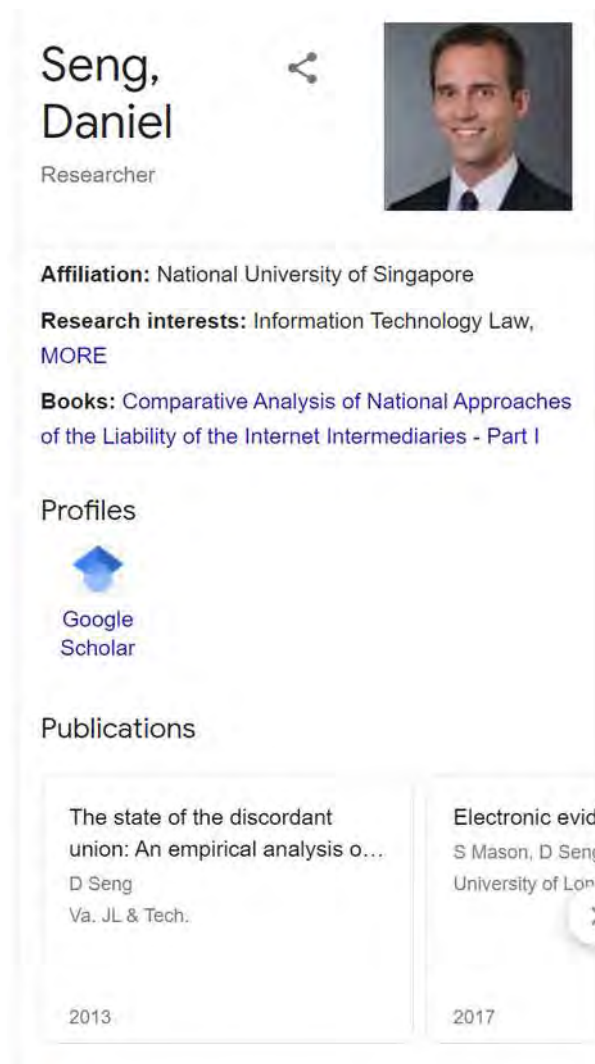


Figure 1: Screen Capture from Google Scholar Researcher Profile page about the author Daniel Seng (circa July 2022)

According to Google, Google Scholar uses “automated software, known as ‘robots’ or ‘crawlers’, to fetch your files for inclusion in the search results. It operates similarly to regular Google search.”⁹¹ It also uses “automated software, known as ‘parsers’, to identify bibliographic data of your papers, as well as references between the papers”.⁹² Google Search in turn runs on AI systems and machine learning models.⁹³ So this profile page generated by Google qualifies as AI evidence.

The information presented by the evidence is, however, a mix of hearsay and real evidence. Disregarding the researcher’s name, which is the supplied input, the academic affiliation, research interests and publications can be either hearsay or real evidence. Where these items of information are culled from websites, they are supplied

⁹¹ Google Scholar, Inclusion Guidelines for Webmasters, <https://scholar.google.com/intl/en/scholar/inclusion.html#crawl>.

⁹² Google Scholar, Inclusion Guidelines for Webmasters: Indexing Guidelines, <https://scholar.google.com/intl/en/scholar/inclusion.html#indexing>.

⁹³ Google, How AI powers great search results, <https://blog.google/products/search/how-ai-powers-great-search-results/>.

testimonially by unknown human agents responsible for maintaining these websites. The author knows because *he* supplied many of these items of information on his university's academic profile page, including his research interests and publication information. However, to the extent that the bibliographic data of some of his publications came from online publication sources such as the Social Science Research Network (SSRN), this information would be real evidence. Interestingly, the displayed photo of the researcher is NOT that of the researcher. It is clearly processed from diverse online sources of photos of images of other researchers with the same name, which the Google AI search algorithm attempts to match to the other information. So, while the presented image of the researcher is processed information, it is derived from photos of other researchers that were provided testimonially by the other researchers. The presented image is clearly a Category 3 evidence that has both real evidence and testimonial characteristics.

In this regard, it is useful to heed the wisdom of Wigmore when he described the assessment process for real evidence, and how this is different from hearsay evidence:

It follows, on the one hand, that autoptic proference [also referred to as real evidence], or the tribunal's self-inspection, is to be distinguished from the use of testimonial and circumstantial evidence as the basis of an inference. *Autoptic preference calls for no inference from the thing perceived to some other thing*; and in this sense, but in this sense only, autoptic preference is not evidence, i.e., not evidence in so far as evidence implies a process of inference. ... It is thus evidence, in the sense that evidence includes all modes, other than argument, by which a party may lay before the tribunal that which will produce persuasion. It is something more than and different from testimonial or circumstantial evidence... [emphasis added]⁹⁴

As such, it is vital to circumscribe the scope of application of the real evidence exception to the rule against hearsay, bearing in mind that such evidence that "calls for no inference from the thing perceived to some other thing" and is not subject to the same scrutiny as hearsay evidence or circumstantial evidence.

Authentication

This leads squarely to the issue of the trustworthiness of the evidence, which applies regardless of whether it is admissible hearsay evidence or real evidence.⁹⁵ There are two qualities of trustworthy evidence: its reliability and its authenticity. The rule of hearsay assesses the reliability of the evidence by determining if the evidence is capable of representing the facts to which it attests.⁹⁶ Authenticity of the evidence on the other hand means demonstrating that the evidence is genuine – that it is what it

⁹⁴ 2 Wigmore, § 1150, at 1346.

⁹⁵ See *Lim Mong Hong v Public Prosecutor* [2003] SGHC 161, [38] per Yong Pung How CJ ("It is every bit as important that a computer print-out tendered should be reliable whether or not it contains hearsay.")

⁹⁶ *Electronic Evidence* (4th Ed) at para. 7.1.

claims to be,⁹⁷ and that its condition is substantially unchanged.⁹⁸ The authenticity of evidence is a condition precedent to its admissibility⁹⁹ - to ensure that "the item is what the proponent claims it is".¹⁰⁰ Wigmore describes this rule of evidence as "an inherent logical necessity", to ensure that this is "a genuine rule of evidence", to ensure that "a certain group of data is or is not sufficient to send the case to the jury."¹⁰¹ In the Singapore Evidence Act, which is based on the Indian Evidence Act 1872, this rule is described statutorily as follows:

Facts necessary to explain or introduce a fact in issue or relevant fact, or which support or rebut an inference suggested by a fact in issue or relevant fact, or which establish the identity of any thing or person whose identity is relevant, or fix the time or place at which any fact in issue or relevant fact happened or which show the relation of parties by whom any such fact was transacted, are relevant in so far as they are necessary for that purpose.¹⁰²

Authentication is also explicitly provided for in the US Federal Rules of Evidence, s 901 and s 31.1 of the Canada Evidence Act, the latter of which reads:

Any person seeking to admit an electronic document as evidence has the burden of proving its authenticity by evidence capable of supporting a finding that the electronic document is that which it is purported to be.¹⁰³

Because of its characteristics, authenticating electronic evidence can encompass:

- the identity of the electronic evidence (whether this was the correct record for evidential purposes),
- its integrity and immutability (whether it was altered, manipulated or damaged between the time it was created and the time it was tendered in evidence),¹⁰⁴
- its authorship (who created it or caused it to be created),
- its recipient (to whom it was addressed or targeted),
- its identity (what systems and processes went into producing it),
- its chronology, locality and sequence (in what order was it produced and when and where was it produced),
- its accuracy (were its contents correct and reliable),

⁹⁷ *Electronic Evidence* (4th Ed) at para 7.1.

⁹⁸ *McCormick on Evidence* (West Publishing Co, 3rd Ed, 1984) at p 686; John Henry Wigmore, *Evidence in Trials at Common Law* (vol 7, §2129) (revised by James H Chadbourn) (Little, Brown, 1978) at p 709.

⁹⁹ Daniel Seng and Sriram Chakravarthi, Computer Output as Evidence Consultation Paper 133-138, Technology Law Development Group, Singapore Academy of Law 2003, at 96; Daniel Seng, Computer Output as Evidence, [1997] Sing JLS 161-3.

¹⁰⁰ Federal Rules of Evidence, Rule 901(a) (Authenticating or Identifying Evidence).

¹⁰¹ 3 Wigmore, § 2129, at 2887.

¹⁰² Singapore Evidence Act, s 9; Indian Evidence Act, s 9. Similar provisions can be found in the Brunei Evidence act, Chapter 108, s 9; Malaysian Evidence Act, s 9; Sri Lanka Evidence Act, s 9.

¹⁰³ RSC, 1985, c C-5.

¹⁰⁴ See e.g., Singapore Electronic Transactions Act, ss 17, 19(1).

- its operation (whether the system or device was properly used and operated), and
- its security (who has or could have accessed it, and its underlying systems and processes, for legal or for illicit purposes).¹⁰⁵

In relation to AI evidence, the following additional authentication issues may arise:

- its fairness and bias, which require that like cases be treated alike (individual fairness), that two classes be treated similarly (group fairness), that cases be treated equally when illicit attributes such as race and gender are removed (fairness through unawareness), that each demographic class gets the same results (equal outcome and equal opportunity), and that people with similar likelihoods should be accorded the same expected utility from the AI system (equal impact).¹⁰⁶ All these may depend on the training data that the AI system producing the evidence has been trained on, or the particular machine learning algorithms selected for this purpose,
- its safety, accuracy and reliability in the context of edge and corner cases for which the results have been rendered, including whether the AI system has been previously verified and validated, and whether results are well calibrated, legal, ethical and common-sensical, and
- its interpretability – whether the results are trustworthy, humanly comprehensible and can be logically explained or have a rational explanation. Also described as transparency, this requires access to the internals of an AI system, its training data, the sensory information received by the AI system and data regarding the AI system’s own deliberative processes and its inference results. This is particularly pertinent since a human person or expert has to testify as to the operation of the AI system, and its inherent interpretability will greatly facilitate its authentication. However, modern AI systems, like modern software, are seldom the product of a single person. AI applications are made up of various components and frameworks that are developed by third parties, use various training sets and other datasets the bulk of which are sourced from outside the organization, and deploy using tools developed or licensed from others. Having an AI system that is designed with interpretability in mind would thus greatly facilitate the authentication process. An example of this would be the rules requiring autonomous vehicles to retain records of the sensory data captured by the vehicle for accident-investigation purposes.¹⁰⁷

¹⁰⁵ Consultation Paper, at 161-163.

¹⁰⁶ Artificial Intelligence, § 27.3, at 993.

¹⁰⁷ See, e.g., r 17(4) of the Road Traffic (Autonomous Motor Vehicles) Rules 2017 (S 464/2017) (requiring data recorder on autonomous vehicle to be kept for at least three years from date of recording, and to record, among others, the vehicle speed, location, status, operator override history, sensor data and camera or video footage), California Title 13 Division 1, Chapter 1, Art 3.7: Testing of Autonomous Vehicles, § 288.06(a)(6).

Presumption of Reliability

At first sight, it would appear overwhelmingly difficult to authenticate electronic evidence. Conversely, this would appear to make it very easy to invalidate electronic evidence, especially if such evidence is tendered to inculcate the accused. To this end, prosecutors and the bench have relied on the common law presumption that states:

“In the absence of evidence to the contrary, the courts will presume that mechanical instruments were in order at the material time.”¹⁰⁸

With the overwhelming prevalence of electronic evidence used in the courts, Commonwealth jurisprudence has gradually abandoned the statutory requirement that computer systems are “reliable” as a precondition for the admissibility of evidence generated by such systems.¹⁰⁹ Reliance on the presumption allowed courts to elide the authentication requirements of electronic evidence. For instance, when the Intoximeter 3000 machine’s authentication was challenged, the Queen’s Bench Division opined that:

As I have already indicated, the justices made no finding which would permit the inference that the Intoximeter was in any way defective or not in proper working order. *It has to be assumed*, certainly for the purposes of the submission of no case to answer, that it was in proper working order and that the proper procedure was followed.¹¹⁰ (emphasis added)

In part,¹¹¹ because of this presumption as well as the technical hurdle that has to be crossed to effect an effective challenge to the authenticity of electronic evidence, this has led to a gross miscarriage of justice in hundreds of cases in the UK involving prosecutions conducted by the UK Post Office Limited (“POL”). POL had sued its former sub-postmasters and postmistresses (“SPMs”) for fraud, theft and false accounting after POL installed a new computer system in 1999 for managing local post office finances. Known as the Horizon IT system, these prosecutions started when the SPMs

¹⁰⁸ United Kingdom, The Law Commission, *Evidence in Criminal Proceedings: Hearsay and Related Topics* (CP No 138, 1997) at para 13.13. This presumption has been approved by the Divisional Court in *Castle v Cross* [1984] 1 WLR 1372 at 1377B, *per* Stephen Brown LJ. There is also a statutory presumption, for which see *Electronic Evidence* (4th Ed) at paras. This presumption is often also expressed as the Latin maxim *praesumuntur omnia rite esse acta*. A similar presumption can also be found in the U.S. Federal Rules of Evidence, § 901(b)(9) (“Evidence About a Process or System. - The following are examples only — not a complete list — of evidence that satisfies the requirement [of authenticating or identifying an item of evidence]: Evidence describing a process or system and showing that it produces an accurate result.”).

¹⁰⁹ See, e.g., the repeal of ss 35 and 36 of the Singapore Evidence Act (Cap 97, 1997 Rev Ed) and s 69 of the UK Police and Criminal Evidence Act 1984 (c 60), and similar reform with respect to the England and Wales Civil Evidence Act 1995 (c 38).

¹¹⁰ *Castle v. Cross* [1985] 1 All ER 86, 90 *per* Stephen Brown LJ.

¹¹¹ In part, because the other reasons include the poor management culture of the Post Office, the lack of objectivity on the part of the advising solicitors, the management’s interference with the work of the forensic accountants that were appointed to conduct an investigation into the Horizon system and the ability for the developer to remotely alter the branch records without any logs or records. See Wikipedia, British Post Office scandal, https://en.wikipedia.org/wiki/British_Post_Office_scandal#Post_Office_management_structure. At the time of writing, the Statutory inquiry into the entire matter is still ongoing.

reported shortfalls in the accounts of the post offices which they managed as reflected in the Horizon system. When POL prosecuted its SPMs to recover the account discrepancies, the SPM defendants challenged the reliability of the Horizon system. In *R v Seema Misra*,¹¹² for instance, defense counsel sought information about 2½ years' worth of transaction information for the post office managed by the defendant. The Prosecution challenged this request, on the basis that this was not a sufficiently focused disclosure, that it would be "enormously expensive" to provide this material and that any computer problem "should be visible, at least the symptoms of it to the user of the computer [which the defendant] was using ... every single day, probably perhaps hundreds of times a day".¹¹³ Despite the defense counsel making a formal application for disclosure, especially since the defendant's expert had just been shown by the Prosecution's expert an extract of the "known errors log" for the Horizon system (which would have recorded all errors that had taken place on the system),¹¹⁴ the trial judge denied the application, holding that the defense had ample material to test the integrity of the system. The administrative-minded response of the judge is telling:

The time comes when trial management issues require a judge to say there must be no more delay. A vast quantity of material has been disclosed and considered and the defence have ample material I am quite satisfied to test the integrity of the Horizon system.¹¹⁵

The defendant Misra was subsequently found guilty by a jury and convicted, as were many other ex-SPMs.¹¹⁶ The dissatisfied SPMs formed an alliance¹¹⁷ and took civil proceedings against POL. Despite POL's contention that "the Horizon system is and was very robust (being more robust than most comparable systems)",¹¹⁸ after a lengthy trial

¹¹² T20090070, in the Crown Court at Guilford, 2010, before his Honour Judge N A Stewart and a jury. For the transcript of the trial, see (2015) 12 *Digital Evidence and Electronic Signature Law Review* 44, Introduction; Documents Supplement. Tim McCormack, 'The Post Office Horizon System and Seema Misra', (2016) 13 *Digital Evidence and Electronic Signature Law Review* 133; "Remote Chance' Horizon Scandal" *Private Eye* (No 1438, 24 February – 9 March 2017) at p 37 (this article quotes from the transcript of the trial published in the *Digital Evidence and Electronic Signature Law Review*).

¹¹³ Day 1, Monday 11 October 2010, at 5B-5E ((2015)12 *Digital Evidence and Electronic Signature Law Review*, Document Supplement). This observation appears to parallel what is known as the "Tapper Condition", described in Colin Tapper, "Discovery in Modern Times: A Voyage around the Common Law World" (1991) 67 *Chi-Kent L Rev* 217 at 248.

¹¹⁴ Day 1, Monday 11 October 2010, at 16 ((2015)12 *Digital Evidence and Electronic Signature Law Review*, Document Supplement). See also Tim McCormack, "The Trial of Seema Misra" *ProblemswithPOL* (19 January 2020) <https://problemswithpol.wordpress.com/2020/01/>.

¹¹⁵ Case Transcript, *R v. Seema Misra*, T20090070 Day 1 Monday 11 October 2010, at 26D-E, <https://journals.sas.ac.uk/deeslr/article/view/2198/2133>.

¹¹⁶ See T20090070, in the Crown Court at Guilford, 2010, before his Honour Judge N A Stewart and a jury. For the transcript of the trial, see (2015) 12 *Digital Evidence and Electronic Signature Law Review*, Documents Supplement, sentencing: transcript of proceedings 11 November 2010.

¹¹⁷ They formed the Justice for Subpostmasters Alliance: <https://www.jfsa.org.uk/>.

¹¹⁸ *Bates v Post Office Ltd (No 6: Horizon Issues) (Rev 1)* [2019] EWHC 3408 (QB) at [440].

in 2019 undertaken this time with expert forensic analysis of the software,¹¹⁹ Frazer J in *Bates v the Post Office Ltd (No 6: Horizon Issues) (Rev 1)*¹²⁰ unequivocally rejected POL's contention and concluded that there were indeed "a large number of bugs, errors and defects ... far larger [in] number than were admitted by the Post Office at the beginning of the [trial], and a far larger number than ought to have been present in the system if [the Horizon system] were to be considered sufficiently robust such that they were extremely unlikely to be the cause of [account] shortfalls in branches."¹²¹ As a consequence, POL acknowledged that it had "got things wrong in our dealings with a number of these postmasters" and in 2020, the CCRC referred 47 Post Office cases on the abuse of process to the Court of Appeal.¹²² The flood of cases led to a statutory public inquiry in 2020, culminating in the unprecedented¹²³ passage of legislation to provide for the quashing of convictions for offences alleged to have been committed while the Horizon system was in use by the Post Office.¹²⁴ All in, more than 2,750 SPMs had been affected by the scandal, and the final cost of compensation is expected to exceed £1 billion.¹²⁵

The Presumption as an Evidential Presumption

The British POL scandal is a stark reminder of the critical importance of authenticating electronic evidence in a court of law. On the evidence then available to the defense (prior to the intercession of Frazer J with expert forensic analysis of the software), particularly the disclosure of the known error logs, the trial judges should have called POL to authenticate the accuracy of the financial accounting records that formed the basis for all the POL prosecutions. It has been surmised that the prosecution and trial judges have operated thoughtlessly on the presumption that computers and the electronic evidence they produce are reliable, leading to calls for the abolition of the

¹¹⁹ *Bates v Post Office Ltd (No 3)* [2019] EWHC 606 (QB) and *Bates v the Post Office Ltd (No 6: Horizon Issues) (Rev 1)* [2019] EWHC 3408 (QB).

¹²⁰ [2019] EWHC 3408 (QB), see Appendix 2.

¹²¹ *Bates v Post Office Ltd (No 3)* [2019] EWHC 606 (QB) and *Bates v the Post Office Ltd (No 6: Horizon Issues) (Rev 1)* [2019] EWHC 3408 (QB).

¹²² See Criminal Cases Review Commission, "The CCRC Refers Eight More Post Office Cases for Appeal – Bringing Total to 47 So Far", press release (3 June 2020), <https://ccrc.gov.uk/the-ccrc-refers-eight-more-post-office-cases-for-appeal-bringing-total-to-47-so-far/>.

¹²³ It is because of the unprecedented mechanism of Parliament legislating to overturn convictions entered by the courts. Select Committee on the Constitution Post Office (Horizon System) Offences Bill, 6th Report of Session 2023-24, HL Paper 126, 24 May 2024, <https://publications.parliament.uk/pa/ld5804/ldselect/ldconst/126/12602.htm>.

¹²⁴ Post Office (Horizon System) Offences Act 2024 (2024 c. 14); Post Office (Horizon System) Offences (Scotland) Act 2024 (2024 asp 6). See Stephen Mason, The Post Office Horizon IT scandal: convictions quashed but the legal causes remain, Internet for Lawyers Newsletter, 19 June 2024, <https://www.infolaw.co.uk/newsletter/2024/06/the-post-office-horizon-it-scandal-convictions-quashed-but-the-legal-causes-remain/>.

¹²⁵ Wikipedia, British Post Office scandal, https://en.wikipedia.org/wiki/British_Post_Office_scandal.

presumption.¹²⁶ This is not without basis, as the response of the trial judges during the disclosure proceedings suggests a level of skepticism as to the validity of the mounted challenges to electronic evidence, and concerns about managing the ensuing costs and delays arising therefrom. In all fairness to the trial judges, the disclosure exercise does require a balancing of the interest of justice against the runaway costs, delays and introduction of unnecessary complexity in a case.¹²⁷

Nonetheless, it is submitted that the presumption of reliability is not biased against the opponent of electronic evidence. It does not operate to presume that the electronic evidence is authentic and unchallengeable upon being tendered, and it is on this basis that the presumption can be reconciled with the balancing exercise in disclosure. As the UK Law Commission explained:

Where a party sought to rely on the presumption, it would not need to lead evidence that the computer was *working* properly on the occasion in question unless there was evidence that it *may not have been* – in which case the party would have to prove that it was (beyond reasonable doubt in the case of the prosecution, and on the balance of probabilities in the case of the defence).¹²⁸ [emphasis added]

In other words, the presumption of reliability is an evidential presumption, as can be seen from statutory formulations of the presumption. For instance, section 146 of the Australian Commonwealth Evidence Act provides:

If it is reasonably open to find that the device or process is one that, or is of a kind that, if properly used, ordinarily produces that outcome, it is presumed (*unless evidence sufficient to raise doubt about the presumption is adduced*) that, in producing the document or thing on the occasion in question, the device or process produced that outcome.¹²⁹ (emphasis added)

¹²⁶ See e.g., Ladkin et al, The Law Commission presumption concerning the dependability of computer evidence: An invited paper, 14 Mar 2020, (2020) 17 Digital evidence and Electronic Signature Law Review 1, <https://doi.org/10.14296/deeslr.v17i0.5143>; James Christie, The Law Commission and section 69 of the Police and Criminal Evidence Act 1984, 22 Sep 2023, (2023) 20 Digital Evidence and Electronic Signature Law Review 62, <https://doi.org/10.14296/deeslr.v20i.5642>.

¹²⁷ See, e.g., Federal Rules of Civil Procedure, Rule 26(b)(1) (Discovery Scope and Limits); Civil Procedure Rules 1998 (1998 No. 3132 (L. 17)), Rule 31.7(2) (Duty of search: factors relevant in deciding the reasonableness of a search).

¹²⁸ UK Law Commission Report No 245, Evidence in Criminal Proceedings: Hearsay and Related Topics (1997), paras 13.13 – 13.14.

¹²⁹ Australian Commonwealth Evidence Act, s 146(2).

Provisions with similar language can also be found in the New Zealand Evidence Act¹³⁰ and the Singapore Evidence Act.¹³¹

A careful reading will inform the reader that this presumption attempts to strike a balance between the fact that the *omnia rite esse acta* presumption is a general presumption that applies across the board to all devices that are in ordinary use, with the acute awareness that computing devices are prone to error.¹³² When the proponent satisfies the two premises of the presumption (“the device or process is one that ... if properly used, ordinarily produces [the expected outcome]” *and* that the electronic evidence sought to be adduced falls into the class of such electronic records or things ordinarily produced), the presumption is that that device or process “produced [or accurately communicated]” the electronic evidence in question. When the presumption operates, the opponent does not have a *legal burden* to rebut the presumption. Instead, to reverse the operation of this presumption, the opponent only has to show, for instance, that the device producing the electronic record belongs to a well-documented class of unreliable devices, or that erroneous output similar to the record in question has been produced from such devices. In other words, the opponent only has to lead *some* evidence that the computing device *may not have been* operating properly or properly used. Doing so raises doubts about the presumption and requires the proponent to discharge his legal burden of proving the reliability of the electronic record. If the proponent fails to discharge this legal burden, the electronic evidence is inadmissible.¹³³ In this regard, the disclosure rules have to operate in sync with the presumption to empower the opponent with such evidence. This is because only the

¹³⁰ New Zealand Evidence Act 2006, s 137 (Evidence produced by machine, device, or technical process), which reads: “(1) If a party offers evidence that was produced wholly or partly by a machine, device, or technical process (for example, scanning) and the machine, device, or technical process is of a kind that ordinarily does what a party asserts it to have done, it is presumed that on a particular occasion the machine, device, or technical process did what that party asserts it to have done, in the absence of evidence to the contrary. (2) If information or other matter is stored in such a way that it cannot be used by the court unless a machine, device, or technical process is used to display, retrieve, produce, or collate it, a party may offer a document that was or purports to have been displayed, retrieved, or collated by use of the machine, device, or technical process.”

¹³¹ Singapore Evidence act, s 116A(1) (“Unless evidence sufficient to raise doubt about the presumption is adduced, where a device or process is one that, or is of a kind that, if properly used, ordinarily produces or accurately communicates an electronic record, the court is to presume that in producing or communicating that electronic record on the occasion in question, the device or process produced or accurately communicated the electronic record.

Illustration: A seeks to adduce evidence in the form of an electronic record or document produced by an electronic device or process. A proves that the electronic device or process in question is one that, or is of a kind that, if properly used, ordinarily produces that electronic record or document. This is a relevant fact for the court to presume that in producing the electronic record or document on the occasion in question, the electronic device or process produced the electronic record or document which A seeks to adduce.”).

¹³² Peter Bernard Ladkin, Bev Littlewood, Harold Thimbleby and Martyn Thomas, ‘The Law Commission presumption concerning the dependability of computer evidence’, 17 Digital Evidence and Electronic Signature Law Review (2020) 1–14, available at <https://journals.sas.ac.uk/deeslr/article/view/5143/5027>. It is submitted that the authors had misapprehended the Law Commission’s view of the presumption as a *legal* presumption in their critique. But this does not detract from the point that they made, that computing devices are prone to error.

party in possession of the electronic evidence – the proponent of the evidence – has the resources, access rights or knowledge to fully understand the system from which the evidence was extracted.¹³⁴ Furthermore, where the disclosure is objected to on the grounds of confidentiality, trade secrecy or protection of intellectual property rights, especially as regards source codes, secret algorithms and training data, the courts must be ready to solve this problem by making various protective orders to prevent the loss of intellectual property rights by restricting access to the code or data, or restricting access to only approved experts, requiring proceedings to be conducted “in camera”, mandating the redaction of court orders and destroying all pertinent exhibits and notes at the close of proceedings.¹³⁵

The application (and misapplication) of the reliability presumption may be illustrated by the Singapore case of *Telemedia Pacific Group v Credit Agricole*.¹³⁶ In this case, among the issues canvassed was whether the defendant bank could prove that it had sent “transaction advices [sic]” to the plaintiff, who claimed not to have received them. To do so, the defendant tendered in evidence printouts from the electronic records of the bank system that processed, generated and dispatched statements of accounts and transaction advices. These were in the form of print logs (showing the transaction advices generated) and the control logs (showing the number of envelopes processed for each transaction advices generated). The defendant’s mailing department head testified that after reviewing these records, he concluded to the best of his knowledge or understanding that the logs corresponded to the envelopes, that these envelopes with the advices had been picked up by Swiss Post and then dispatched, and thus, that the advices had been sent to the plaintiff. For these reasons, the court concluded that the presumption of reliability¹³⁷ was established since the plaintiff had “not produced any evidence to rebut the presumptions of production or accurate communication of the electronic documents”, aside from asserting that it had not received the transaction advices. This, coupled with the other evidence, led the court to conclude that the transaction advices had been dispatched to the plaintiff.¹³⁸

¹³³ This is explored in more detail in AI and Evidence, 250-254. For an example of a protracted legal battle involving the admissibility of electronic records in the U.K. Post Office Limited ledger systems, see *Bates v the Post Office Ltd (No 6: Horizon Issues)* [2019] EWHC 3408 (QB), in particular, Appendix 2.

¹³⁴ See e.g., Case Transcript, *R v. Seema Misra*, T20090070 Day 1 Monday 11 October 2010, at 7H-8A, 15F-G, 15H-16H, 19G-20A (Crown claimed that disclosure of events log for Horizon system would be too expensive although evidence was in possession of the Crown, Crown expert himself admitted that he had relied on others for this information, and Crown admitted it was itself overwhelmed by the transaction logs). <https://journals.sas.ac.uk/deeslr/article/view/2198/2133>.

¹³⁵ Edward J Imwinkelried, “Computer Source Code: A Source of the Growing Controversy over the Reliability of Automated Forensic Techniques” (2017) 66 DePaul L Rev 97. For a discussion of the US position, see, eg, Christopher M Mislow, “Protecting Source Code From Disclosure During Pretrial Discovery” (1984) 12 Utah Bar J 39.

¹³⁶ [2015] 1 SLR 338.

¹³⁷ Codified in the Singapore Evidence Act, s 116A(1).

¹³⁸ [2015] 1 SLR 338, at [260]-[262].

With respect, the focus on whether the mailing system “both ordinarily produces and accurately communicates information of transactions”, such that the reliability presumption applied,¹³⁹ only has the effect of authenticating the operation of the bank's mailing system that printed out the transaction advices and prepared them for posting. The presumption does not go as far as to address the issue of whether the printed transactions were physically sent out and received by addresses, including the plaintiff, via non-electronic postal services. The plaintiff's dispute that it had *not* received its advices in question went to the issue of the dispatch and receipt of the advices, not the printing of such advices. The focus should therefore be on the Swiss Post collection and dispatch processes, which was not a matter to be established by print and control logs tendered, but by evidence as to the mailing system operated by the bank with Swiss Post.¹⁴⁰ It is unfortunate that the reliance on the reliability presumption has caused its over-reaching application and obscured the true issue at hand.

An Example: AI Evidence on Identification

The same approach applies in relation to AI evidence, and while the authentication rules are the same, their application could be more exacting, depending on the issue, context and situation in which the AI evidence is used. To quote from a leading article on the subject:

The greater the risk of unacceptable adverse consequences, the greater the need to show that the AI technology is unlikely to produce those consequences. Judges, tasked with making the initial determination of admissibility of AI evidence under [U.S. Federal Rules of Evidence] Rule 104(a),¹⁴¹ should be skeptical of admitting AI evidence that has been shown to be accurate by no more than an evidentiary coin toss. They should insist that the proponent of the evidence establish the validity and reliability of the AI to a degree that is commensurate with the risk of the adverse consequences likely to occur if the technology does not perform as claimed. And if the proponent of the evidence fails to do so, then the trial judge should evaluate under Rule 403¹⁴² whether the probative value of AI authenticated by a mere preponderance is substantially outweighed by the danger of unfair prejudice to the adverse party or would confuse or mislead the jury to an unacceptable degree, taking into

¹³⁹ *Id.*, at [255].

¹⁴⁰ For a contrasting opinion where it was held that the bank failed to establish that its mailing and dispatch system had operated properly and therefore the bank had failed to discharge its legal burden that it had even physically dispatched the transaction advice to its client, see *Jiang Ou v EFG Bank AG* [2011] 4 SLR 246; [2011] SGHC 149 at [56].

¹⁴¹ U.S. Federal Rules of Evidence, § 104: “The court must decide any preliminary question about whether ... evidence is admissible.”

¹⁴² U.S. Federal Rules of Evidence, § 403: “The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence”.

consideration the nature of the adverse consequences that could occur if the AI technology is insufficiently accurate or reliable.¹⁴³

For instance, when the prosecution's case against the accused is wholly or substantially based on disputed identification evidence, the Turnbull guidelines have to be followed by the judge.¹⁴⁴ This requires warning the jury to take care before relying upon identification evidence, instructing them to look closely at the circumstances in which the observation or listening took place, pointing out any inherent weaknesses in the identification evidence, identifying the quality of the identification evidence (which includes withdrawing the case from the jury if the judge assesses the quality of the identification evidence to be poor) and there is other evidence to support the identification evidence. As the court in *R v Bentley* warned:

The object of the Turnbull warning is to alert the jury to the dangers or evidential pitfalls which might exist in the case before them. Recognition evidence cannot be regarded as free from danger; people may be mistaken even in recognition. There are degrees of dangers: how long had the witness known him, etc. It is advisable to alert the jury to the possibility of mistake.¹⁴⁵

Because of the dangers of using identification evidence wholly or substantially to identify the accused, modern law enforcement and courts have developed standards regulating eyewitness identification evidence.¹⁴⁶ These rules serve to authenticate

¹⁴³ Paul W. Grimm, Maura R. Grossman, and Gordon V. Cormack, *Artificial Intelligence as Evidence*, 19 *Nw. J. Tech. & Intell. Prop.* 9, 94-95 (2021). The authors went on to propose that where AI evidence is adduced, the Daubert factors could be borrowed to assess whether AI evidence should be admitted. "To authenticate AI technology, its proponent must show that it produces accurate, that is to say valid, results. And it must perform reliably, meaning that it consistently produces accurate results when applied in similar circumstances. When the accuracy and reliability of technical evidence has been verified through independent testing and evaluation of the AI system that produced it, the methodology used to develop the evidence has been published and subject to review by others in the same field of science or technology, when the error rate associated with the AI system use is not unacceptably high, when the standard testing methods and protocols have been followed, and when the methodology used is generally accepted within the field of similar scientists or technologists, then it has been authenticated. It does what its proponents say it does. And introducing it produces none of the adverse consequences that Rule 403 is designed to guard against."

¹⁴⁴ *R v Turnbull* [1977] QB 224. The leading case in the U.S. is *Manson v. Brathwaite*, 432 U.S. 98 (1977) (instructing judges to examine whether the identification procedures were unnecessarily suggestive, and to assess whether an identification is reliable by examining: (1) the opportunity of the witness to view the criminal at the time of the crime; (2) the witness' degree of attention; (3) the accuracy of the witness' prior description of the criminal; (4) the level of certainty demonstrated at the confrontation; and (5) the time between the crime and the identification procedure). The U.S. Supreme Court emphasized that "reliability is the linchpin in determining the admissibility of identification testimony." at 114. Similar rules exist in other jurisdictions e.g. Australia Commonwealth Evidence act, part 3.9, ss 113-116; New Zealand Evidence Act 2006, ss 45-46A; *Heng Aik Ren Thomas v Public Prosecutor* [1998] 3 SLR(R) 142, [33]-[35] (Singapore Court of Appeal) (reformulated into a three-step test).

¹⁴⁵ (1994) 99 Cr App R 342; [1991] Crim LR 620.

¹⁴⁶ E.g., the UK Crown Prosecution Service has developed a series of rules that police have to observe when a suspect disputes being the person identified by a witness. Crown Prosecution Service, *Charging (The Director's Guidance)* - sixth edition, December 2020, incorporating the National File Standard: Part A – A suspect disputes being the person identified by a witness (31 Dec 2020), <https://www.cps.gov.uk/legal->

identification evidence, by ensuring that the reliability of the identification evidence can be assessed in a court of law.

Where the eyewitness evidence is replaced with AI identification evidence, the considerations about reliability of such evidence remain, albeit heightened by the complexities of AI, but also fragmented by the different applications of AI based identification. For instance, where AI facial *detection* is used to find a face in an image, its use focuses on finding faces, not on identification of those faces. Where the technology is used for *analysis*, it can be used as a fun tool such as face filters, a tool to validate access e.g. to a mobile phone such as Apple's Face ID, or group related photos together such as photos of family members versus those of friends and office colleagues. Finally, where the facial detection technology is used for *recognition*, this seeks to confirm that the face is correctly identified from a database of individuals.¹⁴⁷ In each of these different contexts,¹⁴⁸ the level of authentication required to demonstrate the reliability of AI facial identification evidence will vary, depending on its respective application. Expert evidence may have a useful role to play to guide the court in assisting the court in assessing the level of authentication required.¹⁴⁹

To illustrate, we can compare the use of AI facial identification evidence for recognition with eyewitness identification. Just as people can be mistaken even in recognition, AI systems can also be "mistaken" in two ways: via false positives, which is to wrongly identify an innocent person as being the suspect, or via false negatives, which is to fail to identify the guilty person as the suspect.¹⁵⁰ These can arise because of the many variations in the in situ image data, such as lighting conditions and exposures,¹⁵¹ varied

[guidance/charging-directors-guidance-sixth-edition-december-2020-incorporating-national-file](#). These include providing information such as whether or not there are any material discrepancies between the witness's given description of the suspect and the actual appearance of the accused, the lapse of time between the original observation and subsequent identification, a statement of the witness as to how long the witness has known the suspect, the nature of their relationship, how often the witness has seen the suspect and in what circumstances and so on, details of the first identification etc.

¹⁴⁷ Thorin Klosowski, Facial Recognition Is Everywhere. Here's What We Can Do About It., NYT, 15 July 2020, <https://www.nytimes.com/wirecutter/blog/how-facial-recognition-works/>.

¹⁴⁸ Google AI, Our approach to facial recognition, <https://ai.google/responsibility/facial-recognition/>.

¹⁴⁹ Paul W. Grimm, Maura R. Grossman, and Gordon V. Cormack, Artificial Intelligence as Evidence, 19 Nw. J. Tech. & Intell. Prop. 9, 95 (2021). "To authenticate AI technology, its proponent must show that it produces accurate, that is to say valid, results. And it must perform reliably, meaning that it consistently produces accurate results when applied in similar circumstances. When the accuracy and reliability of technical evidence has been verified through independent testing and evaluation of the AI system that produced it, the methodology used to develop the evidence has been published and subject to review by others in the same field of science or technology, when the error rate associated with the AI system use is not unacceptably high, when the standard testing methods and protocols have been followed, and when the methodology used is generally accepted within the field of similar scientists or technologists, then it has been authenticated. It does what its proponents say it does. And introducing it produces none of the adverse consequences that Rule 403 is designed to guard against."

¹⁵⁰ Klosowski, Facial Recognition Is Everywhere.

¹⁵¹ See e.g., Tran, C. K., Tseng, C. D., Chang, L., & Lee, T. F. (2019). Face recognition under varying lighting conditions: improving the recognition accuracy for local descriptors based on weber-face followed by

backgrounds, facial expressions, attires, glasses, masks,¹⁵² beards and makeup,¹⁵³ all of which have to be compared against the sterile database of training data in the form of photos taken for government databases such as criminal records or drivers' licenses. Additionally, the training data could be skewed by way of its demographic distribution,¹⁵⁴ which makes it more likely (or less) to identify racial minorities¹⁵⁵ or the female gender.¹⁵⁶ And while we can cross-examine eye witnesses and query them for the basis of their identification and their confidence in their identification, it is not possible to effect the same authentication for facial recognition software, which are "[not] currently available for public audit and are often based on closed-box proprietary systems that researchers cannot investigate."¹⁵⁷ There is thus no way to evaluate the accuracy of such systems for their inherent weaknesses and evaluate the identification quality, in a manner comparable to the Turnbull guidelines.

There is also the additional issue of individual privacy, in whether the meaningful consent of individuals has been secured for the in situ capture of images in public, and the use of their images in training datasets, many of which appear to be drawn from social media sources. For all these reasons, several US states have banned the use of facial recognition for criminal investigation purposes,¹⁵⁸ and in the EU in its AI Act, it is

difference of Gaussians. *Journal of the Chinese Institute of Engineers*, 42(7), 593–601.
<https://doi.org/10.1080/02533839.2019.1644199>.

¹⁵² See Jennifer Valentino-DeVries, *How the Police Use Facial Recognition, and Where It Falls Short*, NYT, 12 Jan. 2020, <https://www.nytimes.com/2020/01/12/technology/facial-recognitionpolice.html>. See also NIST, *Face Recognition Software Shows Improvement in Recognizing Masked Faces* (1 Dec 2020), <https://www.nist.gov/news-events/news/2020/12/face-recognition-software-shows-improvement-recognizing-masked-faces>.

¹⁵³ See e.g., Nitzan Guetta and Asaf Shabtai and Inderjeet Singh and Satoru Momiyama and Yuval Elovici, *Dodging Attack Using Carefully Crafted Natural Makeup* (Sep 2021), <https://arxiv.org/abs/2109.06467>; Kagan Ozturk and Grace Bezold and Aman Bhatta and Haiyu Wu and Kevin Bowyer, *Beard Segmentation and Recognition Bias* (Aug 2023), <https://arxiv.org/abs/2308.15740>.

¹⁵⁴ See e.g., Jake Parker, *What Science Really Says About Facial Recognition Accuracy and Bias Concerns*, 23 Jul 2022, <https://www.securityindustry.org/2022/07/23/what-science-really-says-about-facial-recognition-accuracy-and-bias-concerns/> (claiming that AI facial recognition accuracy is high and that errors were caused by training data tainting).

¹⁵⁵ Alex Najibi, *Racial Discrimination in Face Recognition Technology*, <https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/>.

¹⁵⁶ See e.g., Gender Shades, <http://gendershades.org/overview.html> (reviewing facial recognition software by Microsoft, Face++ and IBM, and finding male subjects accuracy to be higher than female subjects accuracy, lighter subjects accuracy higher than darker subjects accuracy, and that these differences can be as large as 34.4%), which summarizes the research done by Joy Buolamwini and Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, *Proceedings of Machine Learning Research* 81:1–15, 2018, <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.

¹⁵⁷ Klosowski, *Facial Recognition Is Everywhere*.

¹⁵⁸ See Jennifer Valentino-DeVries, *How the Police Use Facial Recognition, and Where It Falls Short*, NYT, 12 Jan. 2020, <https://www.nytimes.com/2020/01/12/technology/facial-recognitionpolice.html>.

regarded as a prohibited AI practice, whose use is limited only to strictly necessary objectives.¹⁵⁹

Computer-Generated Evidence

Given the mutability of electronic evidence, the issue of digitally-manipulated electronic records has always existed since the advent of the computer. However, the rise of AI and the use of generative models has both engendered the flood of “fake” or “digitally-manipulated” electronic information and raised public concerns about this issue. For instance, “digitally-manipulated” photographs or videos, known as “deep fakes”, involve the alteration of existing images or videos in which a person’s image or likeness in that image or video is replaced with someone else’s likeness.¹⁶⁰ These modifications, made possible by using ML and AI techniques such as autoencoders¹⁶¹ and generative adversarial networks,¹⁶² have a high potential to deceive because of their success, including altering the very subtle gestures and movements of a person, particularly of the mouth. Some techniques can even replace the audio streams of a video with synthetic speech that replicates the speech patterns of a well-known person.¹⁶³ Abuses of this technology include misrepresentation of elected leaders and representatives and fake political propaganda, thereby fomenting mistrust and compromising national security.¹⁶⁴ It has been used to feature pornography of innocent people.¹⁶⁵ It has

¹⁵⁹ These objectives are for targeted search of specific victims of abduction, human trafficking, sexual exploitation, searches for missing persons, prevention of terrorist attacks and identification of persons suspected of committed serious criminal offences. Regulation (EU) 2024/... of the European Parliament and of the council, laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 19 Apr. 2024, Art. 5.1(h).

¹⁶⁰ Wikipedia, “Deepfake”, <https://en.wikipedia.org/wiki/Deepfake>. For a high-level introduction, with a brief discussion of the techniques that can be used, see Britts Paris & Joan Donovan, *Deepfakes and Cheap Fakes The Manipulation of Audio and Visual Evidence* (Data & Society, 2019), https://datasociety.net/wp-content/uploads/2019/09/DataSociety_Deepfakes_Cheap_Fakes.pdf.

¹⁶¹ Wikipedia, “Autoencoder”, <https://en.wikipedia.org/wiki/Autoencoder>.

¹⁶² 1Wikipedia, “Generative Adversarial Network”, https://en.wikipedia.org/wiki/Generative_adversarial_network. The method uses two competing neural networks, known as the generator and the discriminator. The generator produces a convincing image to replace the original image, and the discriminator network compares the false footage with the real videos to find flaws. Each item of software learns from the others until the flaws are removed.

¹⁶³ In 2019, criminals used software to impersonate the voice of a chief executive voice to authorize a fraudulent transfer of €220,000, Catherine Stupp, “Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case” *The Wall Street Journal* (30 August 2019), <https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402>. This scam has become more prevalent. See Joe Hernandez, That panicky call from a relative? It could be a thief using a voice clone, FTC warns, 22 Mar 2023, <https://www.npr.org/2023/03/22/1165448073/voice-clones-ai-scams-ftc>.

¹⁶⁴ See, e.g., Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, Michael Tomz, How persuasive is AI-generated propaganda?, *PNAS Nexus*, Volume 3, Issue 2, February 2024, 34, <https://doi.org/10.1093/pnasnexus/pgae034>.

¹⁶⁵ See, e.g., Janko Roettgers, “Porn Producers Offer to Help Hollywood Take Down Deepfake Videos” *Variety* (21 February 2018) <https://variety.com/2018/digital/news/deepfakes-porn-adult-industry->

already occurred in the biomedical sector wherein scientific images and 3D imagery have been falsified through digital manipulation to support fraudulent papers.¹⁶⁶

Various initiatives have been started to combat this problem, ranging from tools to detect digitally-manipulated images and voices¹⁶⁷ to digital watermarking of AI-generated images and videos,¹⁶⁸ to making sexually-explicit images and voices made without consent illegal.¹⁶⁹ However, as the technology supporting deep fakes is also improving,¹⁷⁰ there may come a time when it will not be possible to detect the manipulation,¹⁷¹ although there are attempts being made to counter this problem.¹⁷²

Based on current technologies, detecting a manipulated image takes time, resources, expertise, and the use of appropriate tools. Should the authenticity of a digital image

[1202705749/](https://www.fbi.gov/media/5230605); FBI, Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes, 5 Jun 2023, <https://www.ic3.gov/Media/Y2023/PSA230605>.

¹⁶⁶ By way of example, see Douglas W. Cromeey, 'Digital Images Are Data: And Should Be Treated as Such', *Methods Mol Biol.* 2013; 931: 1–27, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4210356/>; Yisroel Mirsky, Tom Mahler, Ilan Shelef and Yuval Elovici, 'CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning', 28th USENIX Security Symposium (USENIX Security, 2019), https://www.usenix.org/system/files/sec19-mirsky_0.pdf.

¹⁶⁷ This article mentions some of the projects, past and present: Eliza Strickland, "Facebook AI Launches Its Deepfake Detection Challenge At the NeurIPS conference", *IEEE Spectrum* (11 December 2019) <https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/facebook-ai-launches-its-deepfake-detection-challenge>; Yue Wu, Wael AbdAlmageed & Premkumar Natarajan, "ManTra-Net: Manipulation Tracing Network For Detection And Localization of Image Forgeries With Anomalous Features" (2019) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 9543 http://openaccess.thecvf.com/content_CVPR_2019/html/Wu_ManTra-Net_Manipulation_Tracing_Network_for_Detection_and_Localization_of_Image_CVPR_2019_paper.html.

¹⁶⁸ See e.g., Meta, Stable Signature: A new method for watermarking images created by open source generative Ai, 6 Oct. 2023, <https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/>.

¹⁶⁹ See e.g., California Assembly Bill 602 (creating a private cause of action against a person who creates sexually explicit deepfakes and without consent), Bill 730 (prohibits the distribution of deepfake audio or visual media of a candidate to influence political campaigns within 60 days of an election). See National Conference of State Legislatures, Deceptive Audio or Visual Media ('Deepfakes') 2024 Legislation, 7 May 2024, <https://www.ncsl.org/technology-and-communication/deceptive-audio-or-visual-media-deepfakes-2024-legislation> Australia has introduced the proposed Criminal Code Amendment (Deepfake Sexual Material) Bill 2024 that bans the sharing of non-consensual deepfake sexually explicit material.

¹⁷⁰ For an overview, see Henry Ajder *et al*, *The State of Deepfakes Landscape, Threats, and Impact* (Deeprace, September 2019) https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

¹⁷¹ Tero Karras, Samuli Laine & Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks" (2019) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 4401 https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

¹⁷² For which see the US Defense Advanced Research Projects Agency, which has initiated a new project called Semantic Forensics (SemaFor) <https://www.darpa.mil/program/semantic-forensics>; also see Yuezun Li, Ming-Ching Chang & Siwei Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking" 2018 IEEE International Workshop on Information Forensics and Security <https://ieeexplore.ieee.org/document/8630787>; and Scott McCloskey & Michael Albright, "Detecting GAN-generated Imagery using Color Cues" (2019) <https://arxiv.org/pdf/1812.08247.pdf>.

be doubted,¹⁷³ a digital evidence professional will, among others, conduct an investigation based on a reverse image search,¹⁷⁴ analyze the image metadata about the camera and image to detect any inconsistencies in the metadata and the image presented,¹⁷⁵ compute the image exposure including the adjustments that software and cameras will make to the image,¹⁷⁶ determine whether flash was used and analyze the light patterns, shadows and reflections,¹⁷⁷ and analyze of vanishing lines, geometry of shadows, reflection analysis and lens flare. Other techniques that require intermediate technical skills include analyzing the compression schemes for 2D images,¹⁷⁸ reconstructing faces and scenes in 3D, resampling,¹⁷⁹ and examining images for noise.¹⁸⁰

Of course, one should not dispel the possibility that an image that appears to be incredible may in fact be a credible image.¹⁸¹

In a sense, therefore, the evidential treatment of the issue of manipulated digital data is no different from that of other electronic evidence that needs authentication. Issues as to authentication, including issues with digital data that is manipulated, requires the courts to develop a set of clear procedures for managing authentication issues, a healthy appreciation of the limits of the presumption of reliability, as discussed above, and a robust approach towards disclosure or discovery. Only then can these issues be effectively elucidated.

¹⁷³ Hany Farid, *Fake Photos* (The MIT Press, 2019). The following process is described herein.

¹⁷⁴ Although whether this is possible will depend on the sensitivity of the image under consideration and whether the image can be located on the Internet. See Andreas Rössler *et al*, "FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces" (2018), <https://arxiv.org/pdf/1803.09179.pdf>.

¹⁷⁵ Metadata would include data about the camera (make, model), shutter speed, aperture size, focal length, image format, compression, compatibility, geo-location information, date, time, and location tags. The metadata can be used to match an image to a particular device. In addition, when an image is saved and manipulated, the metadata might be modified, augmented, or removed.

¹⁷⁶ This explores the quantitative relationship between the camera settings and the properties of the image: exposure, depth of field, motion blur, ISO (International Organization for Standardization) settings.

¹⁷⁷ For deep fakes, this includes reviewing light patterns on the surfaces of eyes and ears.

¹⁷⁸ This involves looking for artefacts from the JPEG compression scheme such as JPEG signatures and ghosts. See Shruti Agarwal & Hany Farid, A JPEG Corner Artifact from Directed Rounding of DCT Coefficients (TR2018-838) *Dartmouth Digital Commons* (1 February 2018)), https://digitalcommons.dartmouth.edu/cgi/viewcontent.cgi?article=1377&context=cs_tr; Shruti Agarwal & Hany Farid, Photo Forensics from JPEG Dimples, <https://farid.berkeley.edu/downloads/publications/wifs17.pdf>; and Hany Farid, "Image Forensics" (2019) 5 Annual Review of Vision Science 549.

¹⁷⁹ This involves comparing pixel patterns in the image that result from how the image was recorded and manipulation.

¹⁸⁰ This involves assessing imperfections introduced by the software in a camera.

¹⁸¹ Hany Farid illustrates this in *Fake Photos* (The MIT Press, 2019) at pp 38–45 and points out that it can be important to establish whether a gruesome image of a beheading was plausible.

Conclusion

In his Commentaries, Blackstone states:

Evidence signifies that which makes clear or ascertains the truth of the very fact or point in issue, either on the one side or the other.¹⁸²

This reminder could not be more apt in the age of AI and electronic evidence. In the age of the fourth industrial revolution, the need to guard against erroneous persuasion is even more pressing, especially since AI is automating what used to be manual tasks and is used in processes for rendering decisions in ways that will impact human activity. In the absence of any ability to meaningfully cross-examine AI evidence in the absence of its human progenitor, the other rules of evidence, in particular, hearsay and authentication, have to step up to fill the void. It is hoped that this essay demonstrates that there is much which the evidential rules of hearsay and authentication can help to ameliorate these problems. In my opinion, it is precisely because the rules of evidence are forged on the anvils of the adversarial process, hammered in the heat of challenges to evidential trustworthiness and quenched by the need for fairness to the parties, that even though its rules have been crafted way before the advent of AI, evidence rules that are properly applied remain the timeless bulwark against erroneous, unreasonable and unreliable AI evidence.

¹⁸² Sir William Blackstone, Commentaries on the Laws of England, III (1871), 367.