EVALUATING THE EVIDENTIAL VALUE OF EVIDENCE GENERATED BY AI

Preeti Kushwah, O.P. Jindal Law School

Introduction

Our world is becoming increasingly automated. We are surrounded by devices that are designed to simplify our lives. We use search engines to obtain whatever information we require, rely on Google Maps to get us almost anywhere, ask voice-activated assistants like Siri and Alexa for assistance in a variety of ways, etc. With AI invading every aspect of our lives, it is unavoidable that it will have an impact on the legal system as well. The question of how the current legal system would evaluate AI-generated evidence arises from the interaction between AI technology, declining human intervention, and the limitations of the existing legal framework designed only for humans.

Machine learning is the main foundation of the majority of AI systems. Instead of having computers follow pre-programmed rules to carry out complicated tasks, ML works by "enabling computers to learn directly from examples, data, and experience."[1] Because AI differs from typical computer systems in that it does not have pre-programmed rules and is dynamic, it will be treated differently from how a court would evaluate traditional computer-generated evidence. Artificial Intelligence comes with a language problem, so one must start by defining what they mean when they say AI. AI in this paper refers to "software systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal."[2] The idea of an agent

[1] The Royal Society*, Machine learning: the power and promise of computers that learn by example*, (Apr. 2017), https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf.

[2] Eftychia Bampasika, *Artificial Intelligence as Evidence in Criminal Trial*, In CEUR Workshop Proceedings (pp. 133-138). http://ceur-ws.org/Vol-2844/ethics7.pdf.

capable of making decisions with a fair amount of autonomy based on how it perceives and understands its surroundings is fundamental to the idea of artificial intelligence, as described above.

The use of AI in the justice system may prove to be strategically significant and transformative for investigation, fact-finding, and prevention. The idea of using machines to corroborate scientific findings and expert testimony is not new. There is tolerance for both ignorance and risk when using machines for evidentiary reasons; ignorance of how these systems work and the chance that they might not "get it right every time."[3] By decreasing judicial arbitrary behaviour, systematising the proof process, and enhancing trial effectiveness, AI could work as an adjunct mechanism to help the court in its fact-finding process. Therefore, using AI more frequently during court proceedings may reduce the "whim and caprice" of the judge or jury.[4] A high degree of objectivity and precision might be promised by AI when used in the court system. AI by the virtue of its nature is electronic evidence. Electronic evidence can by its very nature be manipulated or even destroyed, either intentionally or unintentionally. The primary concern is ensuring the validity and reliability of such evidence and the systems that store, process, and analyse data. Because AI is subject to biases and other limitations, its outcome accuracy cannot be trusted.

The myriad of limitations that AI has can be broadly classified into the following categories: 1. Bias, 2. Inexplicability and, 3. Lack of accountability. [5]

1) Bias

A key issue with AI is the issue of bias, which can lead to discriminatory effects that are sometimes intended but more frequently unintended. Bias can enter the AI system at any stage i.e., input, processing, and outcome. Further, how humans interpret these outcomes is also coloured with biases.

Machine-learning algorithms learn to work by a loop of data feeding and feedback. As these algorithms are trained using data that has been previously recorded, they may help in

[3] PW Nutter,' *Machine Learning Evidence: Admissibility and Weight'* 21(3),Journal of Constitutional Law 919–58, (2019).
[4] Ben Bryant*, Judges are more lenient after taking a break, study finds*, The Guardian, (Apr. 11, 2011, 8:01 PM), https://www.theguardian.com/law/2011/apr/11/judges-lenient-break.
[5] *supra* note 2.

reinforcing the existing bias, that they are designed to prevent. This is possible because the training data used in AI is not representative of the target population. Two of the most well-known examples are when two Black persons were mistakenly classified as gorillas on Google Photo[6] and, while searching for the word 'CEO', on Google Image search only male images appeared.[7] A similar issue was observed, with facial recognition software, which has difficulties correctly identifying Black women's faces compared to light-skinned women, with error rates ranging from 34.8% to 0.8% because they are not well represented in the training set[8]. For this training set, the data set included 83% white people and 77% men. Data may also be differentially noisy, which means that errors are not spread equally among the different sections. This could be due to missing/lack of data for specific groups or an ineffective collection technique that fails to acquire data uniformly. Defendant Loomis claimed that the COMPAS instrument discriminated on the basis of gender since it assessed male and female offenders separately as research has proven that female offenders differ from male offenders.[9]

Data might also be biased because, though an AI system might not explicitly take a protected class label or feature, like race, into account, the data contains substitutes for those labels or features that the algorithm does take into consideration. For instance, the COMPAS tool requests information regarding drug possession and usage arrests.[10] This question, along with many other characteristics such as zip code, education, job, and incarceration, can easily be used as a proxy for race. Considering the fact that black individuals are known to be arrested by police for drug possession and usage much more frequently than White people are.[11] This kind of data produces biased conclusions because it is more representative of police action and

[6] Maggie Zhang, *Google Photos Tags Two African Americans As Gorillas Through Facial Recognition Software*, Forbes, (Jul 1, 2015m 01:42 PM), https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=6826d1fe713d.
[7]Andrew Van Dam, *Searching for images of CEOs or managers? The results almost always show men*, The Washington Post, (Jan. 3, 2019, 7:00 AM), https://www.washingtonpost.com/business/2019/01/03/searching-images-ceos-or-managers-results-almost-always-show-men/.
[8]Larry Hardesty*, Study finds gender and skin-type bias in commercial artificial-intelligence systems*, MIT News, (Feb. 11, 2018), https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212.
[9] State of Wisconsin v Loomis, 2016 WI 68.
[10] Rhys Dipshan, *Risk Assessment Tools Aren't Immune from Systemic Bias. So Why Use Them*, (Jul. 17, 2020, 7:00AM), https://www.law.com/legaltechnews/2020/07/17/risk-assessment-tools-arent-immune-from-systemic-bias-so-why-use-them/?slreturn=20220929161608.
[11] Peter Walker, *Black people twice as likely to be charged with drugs possession – report*, The Guardian, (Aug. 21, 2013, 7:11PM), https://www.theguardian.com/world/2013/aug/21/ethnic-minorities-likely-charged-drug-possession.

social constructs than it is of recidivism risk. Additionally, bias in data is possible since it reflects society's institutionalised racism and gender prejudice.

It shall also be noted that developers might not be adequately qualified or equipped to make these algorithmic design decisions as they do not reflect the diversity of the populations to which the algorithms will be applied and receive little to no ethics training and may thus be indifferent to the unintended repercussions of their decisions.[12] Lawyers, judges, and regulators enter the picture when these decisions have long been made and are opaque or not readily altered. Silent failures brought on by this oversight frequently go unnoticed until they have disastrous effects on the public's perception of the institution. As a result, the burden is then on the attorneys and judges to ensure that the right questions are being raised, whether impact analyses have been conducted, how and by whom the tool was evaluated for bias, and whether the appropriate metrics were gathered and reported.

Finally, bias develops as a result of how AI system output is interpreted by humans. Human beings owing to their nature have implicit or unconscious biases. One such bias is automation bias,[13] the propensity for people to favour outcomes from automated decision-making systems while ignoring or undervaluing opposing evidence produced independently from such systems, even if they are correct, because people think the automated decision-making system is somehow more 'reliable' or 'objective.'

Decisions made by such algorithms are unreliable and result in objectionable outcomes, as well as citizens' unwillingness to accept its use by the courts because the potential of malfunction remains a significant possibility. Eliminating all of the biases that AI is prone to be one of the most difficult hurdles to the successful application of AI in criminal justice.

2) Inexplicability

While AI is similar to traditional software in the sense that data goes in, and conclusions come out. In between, there is a 'black-box' of calculations that not only is occasionally inaccessible to the experts themselves but also too few in the courtroom would understand.[14] With the

[12] Paul W. Grimm, Maura R. Grossman, and Gordon V. Cormack, *Artificial Intelligence as Evidence*, 19 NW. J.TECH. & INTELL. PROP. 9 (2021).
[13]Pa Consulting, *What is automation bias and how can you prevent it?*, (Aug. 08, 2018), https://www.paconsulting.com/insights/what-is-automation-bias-how-to-prevent.
[14] *supra* note 3.

existence of the black box, the problem of transparency and explicability arises, particularly in terms of how it makes decisions, makes predictions, and categorises data. The transparency problem is that many contemporary AI techniques are inherently opaque, making them difficult to understand. It is not always clear how a machine-learning model was developed or how it functions, particularly when it uses deep learning or neural networks, unlike code, which can be checked for errors/bugs. It is inexplicable why an algorithm confuses a linesman's bald head for a football, robbing viewers of the true action in favour of focusing on the linesman[15] or how it misidentified 28 Congressmen for known convicted felons.[16] Since, there are no bug fixes, developers modify the training data, select different characteristics or parameters to put emphasis on and evaluate the output as a way for the AI to work more efficiently. Owing to the inexplicable nature of AI output, it is inevitable to wonder how the defendant will be able to defend herself and refute the evidence it generates. A similar contention was raised by Loomis in the Loomis case, whereby he argued that his due process rights were violated when the Circuit Court relied on COMPAS for his sentencing rather than his right "to be sentenced based upon accurate information, in part because the proprietary nature of COMPAS prevent [ed] him from assessing its accuracy".[17] Loomis also argued that it was impossible to confirm the veracity of the COMPAS evaluation without having access to the methodology used to establish the risk score and how the criteria were weighted.[18] Northpointe, Inc., the developer of COMPAS took up the defence of "a proprietary instrument and a trade secret" and refused to reveal the methodology used to compute the risk scores or how the weighting of the factors was done.[19] Regardless of citing research that questioned the reliability of COMPAS and its propensity to unfairly categorise minority offenders as higher risk due to circumstances that may be beyond their control, the Court decided that the tool may be used with the appropriate precautions.[20]

The interpretability problem, the technical difficulty of explaining AI decisions, has given rise to an entire field of study known as 'Explainable AI (XAI)'. The proponents of XAI contend

[15] Surjit Patowary, Robot Cam confuses linesman's bald head for a football in Scotland, Thick Accent, (Oct. 28, 2020), https://www.thickaccent.com/2020/10/28/robot-cam-confuses-linesmans-bald-head-for-a-football-in-scotland/.
[16]Sean Hollister, Amazon facial recognition mistakenly confused 28 Congressmen with known criminals, CNET, (Jul. 26, 2018, 12:45 PM), https://www.cnet.com/news/privacy/amazon-facial-recognition-thinks-28-congressmen-look-like-known-criminals-at-default-settings/.
[17] *supra* note 12.
[18] *Id*.
[19] *Id*.
[20] *Id*.

that AI cannot be trusted unless it can be explained to humans, given this fact they accept that the degree or kind of explanation may differ for various applications or consumers. Outlined by NIST four XAI guiding concepts are as follows:

I. Explanation - providing supporting evidence or the rationale behind all outcomes
II. Meaningful – provided explanation are legible to individual users
III. Explanation Accuracy - ensuring that the explanations accurately disclose the AI system's method for producing the outputs
IV. Knowledge limit - it has enough confidence in its results.[21]

DARPA or Defense Advanced Research Project Agency is one such transparency XAI initiative that strives to develop "glass-box" models that can be discussed as a "human-in-the-loop" without compromising AI functionality.[22] Glass-box is seen to be a workable solution to the black-box problem. Here, moral principles are translated into explicit, verifiable norms that impose limitations on inputs and outputs, creating a "glass box" around the system.[23] The emphasis on inputs and outputs enables the comparison and verification of a wide range of intelligent systems, from agent-based systems to deep neural networks.[24] The objective is to create value-based explanations by ensuring the system adheres to specific social, ethical, and legal values. Some light must be shed on this black box in order for the bench or jury to make an educated conclusion on the defendant's guilt. In conclusion, establishing an AI system's validity and dependability become more crucial for lawyers and judges when it is not transparent or explainable.

3) Lack of Accountability

Another significant challenge that plagues the algorithm is the problem of accountability. Both qualitatively and quantitatively, the current legal framework for the regulation and governance of AI is deficient. The existing legal system is unprepared to address the unique powers and characteristics of artificial intelligence. For instance, Stephen L. Thaler of the Artificial

[21] Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. *Four principles of explainable artificial intelligence*. Gaithersburg, Maryland. (2020).
[22] Matt Turek, *Explainable Artificial Intelligence (XAI),* Defense Advanced Research Projects Agency, (Oct. 8, 2022, 3:51 PM) https://www.darpa.mil/program/explainable-artificial-intelligence.
[23]Andrea Aler Tuvella, et. Al., *Governance by Glass-Box:Implementing Transparent Moral Bounds for AI Behaviour*, *arXiv preprint arXiv:1905.04994* (2019), https://www.ijcai.org/proceedings/2019/0802.pdf.
[24] Arun Rai*, Explainable AI: fromBlack Box to Glass Box*, 48 J. ACAD. MARKETING SCI. 137 (2020)

Inventor Project tried to patent two inventions under DABUS. His patent application was rejected by U.K. Intellectual Property Office, European Patent Office and U.S. Patent and Trademark Office citing the need for the 'inventor' to be a natural person.[25] Contrastingly South Africa and Australia allowed for the patent under DABUS for its AI-generated invention. There seems to be no common ground as of right now for the treatment of AI and AI-generated matter. Where the algorithms are self-learning and are capable of acting in increasingly unpredictable ways. Who should be held liable for errors and accidents brought on by Artificial intelligence systems, and under what conditions, is not currently specified under the law governing liability for AI systems? There are a variety of potential parties involved: the data analyst or collector, the inventor, the designer or developer, the manufacturer, the retailer, the consumer, the AI itself, a combination of the aforementioned, or none at all.[26] In this instance where U.S. Immigration and Customs Enforcement ('ICE') changed its risk-assessment algorithms that produce only one result would always recommend 'detain' for immigrants in custody,[27] who is to be held accountable? Furthermore, because of the dynamic nature of the technology and its 'black-box' design, it might be hard to figure out how and why the system came to the choice it did or to reverse-engineer the decision-making process.

The lack of resilience in AI is a further problem. The ability of AI systems to identify and withstand deliberate and unintentional attempts to sabotage machine learning models or to otherwise adjust to risk is known as resilience. While companies take measures to protect AI systems from cyber-attacks, with the advancement in technology, in a wicked loop, skilled hackers quickly discover ways to get around these protective measures.

AI algorithms are employed in adversarial machine learning to jeopardise the working of AI systems. It is accomplished by providing misleading data inputs with the intention of triggering a malfunction. These attacks affect any kind of system built on machine learning algorithms and range from spam emails disclosing sensitive data to deepfakes. Use of a telegram AI-bot to create pornographic deepfakes to abuse women in thousands[28] or turn an AI

[25]Ryan Abbott, *The AIP—Now With More Copyright!*,Artificial Inventor, (Jul. 10, 2022), https://artificialinventor.com/867-2/.
[26] *supra* note 12.
[27] Nikhil Sonnad, *US border agents hacked their "risk assessment" system to recommend detention 100% of the time*, Quartz, (Jun. 26, 2018), https://qz.com/1314749/us-border-agents-hacked-their-risk-assessment-system-to-recommend-immigrant-detention-every-time.
[28]Matt Burgers, *A deepfake porn bot is being used to abuse thousands of women*, Wired, (Oct. 20, 2020, 03:00 PM), https://www.wired.co.uk/article/telegram-deepfakes-deepnude-ai.

chatbot into a racist Twitter troll.[29] It is getting more and more difficult to differentiate between AI-generated content and human-generated material. In some cases, fake evidence might be so ingenious and compelling that only a forensics expert could tell if it is authentic or not. However, in most situations, hiring an expert witness can be rather pricy. Genuine seeming fake evidence alongside automation bias may result in faulty decision-making by lawyers, policymakers and judges leading to a loss of trust and confidence in the institution.

All of these unique qualities of AI translate into actual and significant problems with the admission of electronic evidence, whether it takes the form of records created by AI systems or real evidence. We must take into account their reliability, authenticity, and validity when assessing the evidentiary value of AI-generated evidence.

Reliability

The consistency of an AI system's output is referred to as reliability; specifically, whether the same (or a highly correlated) outcome is produced under the same set of conditions.[30] The discussion in the previous section of how complex AI systems is raising the question of their reliability a notch higher. This is particularly true when AI systems contain hidden errors that are triggered under peculiar conditions. The algorithm camera mistook a lineman's bald head for the football and kept focusing on it throughout the game as an illustration of an unreliable AI in action.[31] Although this is a fairly subdued illustration. The instance of Uber's autonomous vehicle killing a pedestrian due to its inability to classify the victim as a human is not.[32] Later investigations found that the system design did not take jaywalking pedestrians into account; this, along with a distracted driver, happened to be the cause of the disaster. [33] It is strange to imagine that an AI-based car designed to function in public settings was not vetted for its environmental perception and did not take jaywalking, which is a widespread transportation behaviour, into account. Therefore, a variety of techniques will need to work together to demonstrate the reliability of AI systems. Every single component of the AI system,

[29] James Vincent, *Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day*, The Verge, (Mar. 24, 2016, 4:13 PM), https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.
[30] Supra note 8.
[31] Supra note 11.
[32]Phil McCausland, *Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk*, NBC News, (Nov. 10, 2019, 1:58AM), https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281.
[33]*Uber in fatal crash had safety flaws say US investigators*, BBC News, (Nov. 6, 2019), https://www.bbc.com/news/business-50312340.

many of which are ML systems in and of themselves, will need to be tested. The testing environment for certifying AI systems will need to evaluate the fundamental accuracy of the system that incorporates all ML subsystems based on well-known settings or parameters that have accurate or known results. How many of these different settings an AI system has been evaluated in and for which it has been certified will be a major determinant as to how reliable it is. Providing feedback input after reviewing any errors the AI may have or any indications that it was unable to effectively manage the different parameters. Based on the information containing the number, rate and nature of these errors, the Court can draw its decision regarding the reliability of such systems and their compliance with the prevailing standards.

Under the Indian law proving reliability of an evidence is a pre-requisite to its admissibility.[34] The reliability standards for evidence are not stated explicitly in Indian law. Judges are tasked with using their judgement, knowledge, and discretion to decide whether or not a piece of evidence is reliable under the law.

Authenticity

For evidence to be trustworthy, it must be reliable and authentic. The quality of authenticity shows how credible the evidence is. For a piece of evidence to be admissible in a court of law, it must be authentic. Because electronic evidence is inherently malleable, its authenticity is a crucial factor in determining its admissibility in the court. A piece of evidence's authenticity serves as the foundation for evaluating its integrity, immutability, authorship, accuracy, and security. To allow and encourage rigorous evaluation of authentication issues, a clear system should be established that specifies how concerns about the legitimacy of electronic evidence should be addressed and treated.

The admissibility of electronic evidence is covered in Section 65B of the Indian Evidence.[35] The non-technical requirements, according to subsection 4 of the section, include the need for an authenticity certificate.[36] A person holding a responsible position in regard to the device via which the data has been produced must execute/sign the certificate. The certificate is required to identify the electronic record containing the statement, explain how it was generated, and provide any information about producing device that may be necessary to demonstrate that the

[34] R.M. Malkani v. State of Maharashtra, (1973) 1 SCC 471.
[35] The Indian Evidence Act, §65B, No. 01, Acts of Parliament, 1872 (India).
[36] The Indian Evidence Act, §65B (4), No. 01, Acts of Parliament, 1872 (India).

electronic record was generated by a computer. According to the Supreme Court's ruling in the case of Anwar PV v. PK Basheer and Others, a certificate under Section 65B of the Evidence Act is mandatory.[37] The primary aim of the certification is to guarantee the reliability of the source of the data and its authenticity so that the Court can rely on it. This is crucial since electronic data is more likely to be altered and tampered with. This mechanism can also be made applicable to evidence produced by AI. The certificate is a relatively low threshold to establish the reliability of a piece of evidence. As long as the authenticity certificate is produced, a court will be compelled to record the evidence even if it has reservations about the validity and reliability of the evidence. The amount of fraudulent and/or irrelevant evidence that parties enter into the record will almost certainly rise if this authentication approach is relied upon. This will undermine the search for the truth and lengthen the trial, adding to systematic delays.

Validity

Validity refers to the quality of being right or accurate, or, more specifically, as to whether and how precisely an AI system measures what it is designed to measure. One of the ways to test for the validity of the AI-generated evidence can be by employing the Daubert Standard as established in the Daubert v. Merrell Dow Pharmaceuticals, Inc. case.[38] The Daubert Test consists of five steps, including determining whether the method was tested, peer-reviewed, or published, the known rate of error, adhering to and maintaining standard procedures, and whether the method is accepted by the scientific community.[39] In the case of AI, when the accuracy and validity of technical evidence have been verified through separate testing and evaluation of the AI system that generated it, the method and data used have been published or evaluated by other experts in the same area, the error rate is not inexcusably high, the relevant procedures and standards have been adhered to, and the technique employed is regarded as authentic by the scientific community. In deciding whether to allow the evidence at all, the court should be guided by the responses to these points. Allowing the use of AI evidence for a particular purpose that has not been shown to be valid and reliable is detrimental. The increased risk of unfairly confusing or misinforming the judge and jury overshadows the importance of the evidence's probative value.

[37] Anvar P. K. vs. P.K Basheer &Ors. (2014) 10 SCC 473.
[38] Daubert v Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
[39] *Id*.

Conclusion

Despite being relatively new, the application of AI is rapidly expanding throughout every significant segment of our society. Because AI is developing so quickly, the law of evidence defining the conditions in which AI technology and its results should be admitted is still in a nascent stage. Placing the burden on lawyers and judges to be vigilant when questioning the admissibility of evidence is a formula for failure. Most lawyers and judges will not be up to the task unless they have at least a basic understanding of what artificial intelligence is, how it works, how to evaluate it scientifically and statistically, and the issues that need to be addressed in order to make decisions about its validity and reliability, and consequently its admissibility. To do this, one must take into account the complex, dynamic nature of AI as well as its constraints. AI evidence should be carefully examined and not just taken at face value.

AI evidence should be carefully scrutinised and not just accepted at face value, while guaranteeing compliance with current laws and regulations, ensuring ethics and robustness, and accounting for technical and social viewpoints.