CHAPMAN LAW REVIEW

Fingerprints of Injustice: The Truth Behind Artificial Intelligence and Algorithmic-Driven Evidence

Aubrey A. Butler

Fingerprints of Injustice: The Truth Behind Artificial Intelligence and Algorithmic-Driven Evidence

Aubrey A. Butler[*]

Artificial intelligence (AI) has steadily grown in prominence, reaching into nearly every aspect of daily life, and the legal system is not immune from its influence. As various forms of machine evidence have been admitted in court, there has been an accompanying rise in concerns from researchers, judges, and defense attorneys as to the trustworthiness and reliability of this new category of evidence.

Though AI and machine evidence takes various forms, this Note focuses on some of the most prominent programs being used in courtrooms today, namely the probabilistic genotyping software TrueAllele and the recidivism rate prediction software COMPAS. The creators of these programs guarantee their accuracy, yet several studies have demonstrated proven defects in their operation—defects which cannot be fully tested and resolved due to the proprietary or "black box" nature of the underlying source code.

Several solutions have been suggested for how to approach machine evidence moving forward, but this Note posits a new mechanism which has not been previously addressed: the creation of a new federal agency focused on AI within the United States with a department wholly dedicated to computer-driven evidence in the legal system. This agency would be able to analyze machine evidence and send out scientific advisors to courts to counsel judges about the potential dangers of this form of evidence in a way that is not possible under the current system.

[*] Aubrey Butler graduated with honors from the University of California, Los Angeles with a Bachelor of Science in biology. She is currently a student at the Dale E. Fowler School of Law at Chapman University in California where she is a recipient of both a scholarship and the Distinguished Student Fellowship. She is also a Managing Editor of the prestigious *Chapman Law Review*. She has written extensively inside and outside the classroom with publications in the genres of both fiction and nonfiction.

I. INTRODUCTION

The world is ever changing with new ways to create an easier and more streamlined existence that can meet society's growing needs; but progress and justice are not always aligned. A 2021 survey showed that thirty-seven percent of Americans are worried about the growth of artificial intelligence (AI), particularly as it relates to replacing jobs,[1] but its implications in the legal field are equally concerning. Predictive algorithms, large language models, and probabilistic genotyping—often collectively referred to as types of AI—have already made their way into the court system in a development one judge described as "the beginning to a disastrous end."[2] This disastrous end being the replacement of human judgment by the supposedly objective assessments of computer programs with proven defects.

Cybergenetics' TrueAllele software has identified defendants via mixed DNA in nearly one thousand criminal cases,[3] predictive risk assessment algorithms have aided judges in assessing individuals in the criminal justice system since 1998,[4] and generative AI has already gained popularity as a tool for legal research.[5] Yet, TrueAllele faces constant scrutiny for hiding its source code, assessment software such as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) was found to be inaccurate twenty-nine percent of the time,[6] and generative AI lacks the necessary safeguards to filter false or misleading information.[7] When life and liberty are at stake, such concerns cannot—and should not—be ignored.

[1] LEE RAINIE ET AL., AI AND HUMAN ENHANCEMENT: AMERICANS' OPENNESS IS TEMPERED BY A RANGE OF CONCERNS 22 (2022), https://www.pewresearch.org/wp-content/uploads/sites/20/2022/03/PS_2022.03.17_AI-HE_REPORT.pdf [https://perma.cc/3KYS-L24J].

[2] Ed Cohen, *Most Judges Haven't Tried ChatGPT, and They Aren't Impressed*, THE NAT'L JUD. COLL. (July 21, 2023), https://www.judges.org/news-and-info/most-judges-havent-tried-chatgpt-and-they-arent-impressed/ [https://perma.cc/96HJ-Y5PV].

[3] Justin Jouvenal, *A Secret Algorithm Is Transforming DNA Evidence. This Defendant Could Be the First to Scrutinize It.*, WASH. POST (July 13, 2021), https://www.washingtonpost.com/local/legal-issues/trueallele-software-dna-courts/2021/07/12/66d27c44-6c9d-11eb-9f80-3d7646ce1bc0_story.html [https://perma.cc/X62G-9JCX].

[4] *See* Julia Dressel & Hany Farid, *The Dangers of Risk Prediction in the Criminal Justice System*, MIT CASE STUD. IN SOC. & ETHICAL RESPS. COMPUTING, Feb. 5, 2021, at 1, 3.

[5] *See* Bernice Bouie Donald et al., *Generative AI and Courts: How Are They Getting Along?*, PLI CHRON., Sept. 2023, at 1, 4.

[6] *See* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/52YN-M7G7].

[7] *See* Tiernan Ray, *Generative AI Can't Find Its Own Errors. Do We Need Better Prompts?*, ZDNET (Oct. 31, 2023, 8:01 AM), https://www.zdnet.com/article/generative-ai-cant-find-its-own-errors-do-we-need-better-prompts/ [https://perma.cc/F4K2-G4MV].

With the pervasiveness of AI and algorithmic evidence, it would be largely impractical and shortsighted to reject their merits entirely, but measures must be put into place to mitigate the inherent risks. At the forefront of the issue is the fact that companies generally refuse to release their proprietary software's coding, leaving attorneys and judges blind as to how the AI reached a certain conclusion or analyzed the raw data.[8] Despite the lack of transparency, courts have consistently allowed this type of black box evidence to be admitted without requesting that companies release their source code for examination.[9] Among these cases are familiar refrains from companies arguing that keeping trade secrets is necessary to protect business[10] and that defense teams would struggle to decipher the complicated source codes even if they were released.[11]

Recognizing the need to address this growing problem, Chief Justice Roberts made AI the focus of his 2023 year-end report.[12] He warned of AI's various shortcomings, including the possibility of embedded biases, potential violations of due process, and a general lack of reliability, especially pertaining to "hallucinations," the phenomenon where generative AI produces fabricated information.[13] "Machines cannot fully replace key actors in court," the Chief Justice proclaimed.[14] "Nuance matters," and "legal determinations often involve gray areas that still require the application of human judgment. . . . AI is based largely on existing information, which can inform but not make such decisions."[15]

Several solutions have been proposed to overcome this growing conflict, and while many champion amending current

[8] *See* Katherine Kwong, *The Algorithm Says You Did It: The Use of Black Box Algorithms to Analyze Complex DNA Evidence*, 31 HARV. J.L. & TECH. 275, 287–88 (2017).

[9] *Id.* at 284–87, 298; Linton Mann III & William T. Russell Jr., *Disclosure of Software Source Code Not Required to Establish Acceptance of DNA Evidence*, LAW.COM: N.Y. L.J. (May 17, 2022, 12:00 PM), https://www.law.com/newyorklawjournal/2022/05/17/disclosure-of-software-source-code-not-required-to-establish-acceptance-of-dna-evidence/ [https://perma.cc/XVC6-DYSM] (discussing the court case that ruled disclosure of TrueAllele source code was unnecessary because the algorithm was generally accepted in the relevant scientific community).

[10] *See* Kwong, *supra* note 8, at 293.

[11] *See* Jouvenal, *supra* note 3.

[12] *See* JOHN G. ROBERTS, JR., 2023 YEAR-END REPORT ON THE FEDERAL JUDICIARY 2, 5–6 (2023), https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf [https://perma.cc/YY3T-V68E].

[13] *Id.*

[14] *Id.* at 6.

[15] *Id.*

admissibility rules to allow access to source codes, even if that programming were released to lawyers, they would likely be unable to understand such highly technical information outside their expertise.[16] Others argue that the current rules are more than adequate to evaluate modern evidence, and at most, there must simply be new ways of interpreting those rules when AI is at issue.[17] But in the end, maintaining the status quo would be ineffectual at solving problems that were not in existence at the time the rules were developed.

Instead, this Note offers a new solution which has yet to be fully addressed in any existing literature: a federal agency dedicated to tackling the rising incidence of AI with a division wholly focused on machine evidence in the legal system. This division would review AI and machine-driven evidence being offered in court on a specific case and send a court-appointed scientific advisor trained on the topic to advise the judge. The advisor would be able to better understand complicated source code, be available to guide judges about issues they are not able to research on their own, and would remain subject to confidentiality, solving worries about proprietary software codes being released. Furthermore, having a centralized agency in charge would allow a more intensive and collaborative vetting process when it comes to AI evidence.

Part II begins this discussion with a brief overview of the history of AI, including the distinctions among terms such as "machine learning" and "large language models," and how algorithmic evidence steadily made its way into the court system. Part III takes a deeper look at the problems underlying this type of evidence, from potential violations of the Confrontation Clause to outright false conclusions. Part IV surveys the myriad of solutions that have been posited regarding AI and critically analyzes why each fails to satisfy every facet of this complicated issue. Finally, Part V gives an in-depth look at this new proposition, addressing both the logistics and the fact that it is not an entirely novel idea, but an existing mechanism that can

[16] *See infra* Section IV.A; *see also* David A. Prange & Benjamen C. Linden, *Explaining the Almost Unexplainable: Preparing and Presenting Source Code Evidence at Trial*, LAW.COM (June 4, 2020, 10:00 AM), https://www.law.com/legaltechnews/2020/06/04/explaining-the-almost-unexplainable-preparing-and-presenting-source-code-evidence-at-trial/ [https://perma.cc/YAK5-HYQK] (explaining that source code evidence presents "a significant challenge" because of its "obtuse and difficult" nature, the "sheer volume of [which] . . . may require retention of a separate expert" by counsel).

[17] *See infra* Section IV.B.

simply be tweaked and applied to the context of AI. Part VI briefly concludes.

II. BACKGROUND

A. The Inception and Limitations of AI

"AI" is an omnipresent fixture of the modern world, two letters blazed across headlines and billboards that have inescapably become part of society's vernacular—two letters with origins far more humble and hopeful than the often sensationalized concerns of "AI takeover"[18] might lead one to believe. In fact, at its inception, AI was never intended to replace human thought or ability, only to enhance the efficiency of fundamentally objective tasks.

Long praised as the father of modern computer science, Alan Turing laid the groundwork for the possibility of AI in 1950 when he wrote a revolutionary paper based upon the simple premise: "Can machines think?"[19] In answering this question, he developed the "imitation game," now known as the Turing Test, which asks whether a blind interrogator evaluating a conversation between a human and a machine could correctly determine which of the two was the person.[20] If the results are indistinguishable and the interrogator cannot correctly choose, the computer can "think," and it passes the test.[21]

At the time Turing published his paper, no machine could come close to winning the imitation game due to a significant limitation in computing ability—"they couldn't store commands, only execute them."[22] A hallmark of human intellect is the ability to learn from mistakes, which is only possible if one remembers they made a mistake in the first place. When a computer cannot store commands, it cannot learn from past behavior, and its intelligence is accordingly restricted.[23] Undeterred by the obstacles surrounding this novel field, computer scientists

[18] Conor Friedersdorf, *Is This the Start of an AI Takeover?*, THE ATLANTIC (Jan. 3, 2023), https://www.theatlantic.com/newsletters/archive/2023/01/is-this-the-start-of-an-ai-takeover/672628/ [https://perma.cc/8748-SQSH].

[19] A.M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433, 433 (1950).

[20] *See id.* at 433–34.

[21] *See id.*

[22] Rockwell Anyoha, *The History of Artificial Intelligence*, HARV. UNIV.: SCI. IN THE NEWS (Aug. 28, 2017), https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ [https://perma.cc/L2ZW-FEKM].

[23] *See id.*

embraced the challenge of creating a thinking machine with fervor, and just five years later, the world's first AI was born.[24]

Stored on punch cards, reliant on the heuristic approach of trial and error, and still debated for its title as "the first AI program," Logic Theorist was programmed to simulate human problem-solving and was able to prove thirty-eight out of the fifty-two mathematical theorems described in *Principia Mathematica*.[25] Logic Theorist was a breakthrough for the time because "it was the first program in symbolic AI, which uses symbols or concepts, rather than data, to train AI to think like a person."[26] As technology advanced and computers became faster and capable of storing greater amounts of data, AI also progressed, and in 1997, IBM's program Deep Blue became the first computer to beat the world's then-reigning chess champion in a match.[27]

As of today, AI has surged into almost every sector through advancements in "natural language processing, image recognition, and automation," and AI adoption by major companies has increased forty-seven percent since 2018.[28] But despite its now overwhelming presence in society, AI is still relatively misunderstood by the general public, and every computer scientist has a different idea of how to define it.[29]

At its heart, AI is a sequence of ones and zeroes.[30] While humans solve problems using abstract thought, machines follow commands written in the only language they can read—binary. The process of doing so is fairly straightforward: a programmer writes an instruction using a programming language, that instruction is translated into binary code (also called machine

[24] *See id.*

[25] Sarah Sloat, *The First AI Started a 70-Year Debate*, POPULAR SCI. (Oct. 3, 2023), https://www.popsci.com/technology/the-first-ai-logic-theorist/ [https://perma.cc/S9DC-FYVY].

[26] *Id.*

[27] Anyoha, *supra* note 22. However, while Deep Blue managed to win the match, its victory was a result of its speed, not its smarts, as no computer has ever been able to exhibit "humanlike intelligence." *See* Eric Holloway, *For Computers, Smart Is Not the Same Thing as Fast*, MIND MATTERS (Mar. 23, 2021), https://mindmatters.ai/2021/03/for-computers-smart-is-not-the-same-thing-as-fast/ [https://perma.cc/4TW4-7N6X].

[28] *The Current Status of Artificial Intelligence*, ALLTECH MAG., https://alltechmagazine.com/what-is/current-status-of-artificial-intelligence/ [https://perma.cc/XA58-VJL3] (July 29, 2024, 12:47 PM).

[29] *See id.*

[30] *See* Ian Buckley, *What Is Coding and How Does It Work?*, MUO, https://www.makeuseof.com/tag/what-is-coding/ [https://perma.cc/Z5Q4-TPKQ] (June 3, 2021).

code), and the computer follows the instruction.[31] Coding is simply a way of telling the computer what to do; the more complex the command, the more lines of code.

Despite its complicated-sounding name, "algorithm" is just a broad term used to describe a set of instructions for solving a particular problem or executing a particular command.[32] Initial data is input into the algorithm, filtered through the set of directions—which generally take the form of mathematical formulas and problem-solving processes—and final output data is expressed.[33] There are multiple types of algorithms for different applications, but a common, recognizable example is search algorithms, which take input data in the form of key words, search relevant databases for those words, and return the results.[34] Because of the ability to adapt algorithms to perform a variety of functions, they remain vital building blocks of software programs and AI, no matter which form they take.

Terms such as machine learning (ML) and large language model (LLM) are sometimes used interchangeably with AI, but each represents distinct ideas in computer science, and each has inherently different risks and drawbacks. ML represents a subset of AI which uses data and algorithms "to imitate the way that humans learn," gradually improving its accuracy.[35] Unlike algorithms that are used to execute simple commands, ML algorithms are designed to make predictions about patterns of data, determine the error rate of that prediction, and use a model optimization process to better the program's ability to correctly predict outcomes.[36]

ML algorithms are generally trained by inputting data that already have a known output to judge their predictive accuracy, and in this way, the program "uses statistical techniques to help it 'learn' how to get progressively better at a task, without necessarily having been programmed for that certain task."[37] ML

[31] *Id.*

[32] Alexander S. Gillis, *Definition: What Is an Algorithm?*, TECHTARGET, https://www.techtarget.com/whatis/definition/algorithm [https://perma.cc/3WU3-3DES] (last visited Feb. 18, 2025).

[33] *See id.*

[34] *See id.*

[35] *What Is Machine Learning?*, IBM, https://www.ibm.com/topics/machine-learning [https://perma.cc/64P2-RH2P] (last visited May 4, 2025).

[36] *Id.*

[37] Ellen Glover, *What Is Artificial Intelligence (AI)?*, BUILT IN, https://builtin.com/artificial-intelligence [https://perma.cc/KS4R-98F3] (Dec. 3, 2024).

differs from traditional programming in that it is capable of solving more complex problems, such as recognizing faces or making predictions, and it is trained on sets of data rather than simply running code line by line.[38] However, ML systems still involve "input data . . . fed to an algorithm" just as traditional programming involves algorithms built on top of one another.[39] This means ML accuracy is partly reliant on the initial code or model used to teach the software how to recognize patterns, even if its output is ultimately less predictable than with traditional programming.[40] What actually qualifies as good accuracy for ML programs is highly subjective, and the industry standard is set at a success rate of seventy percent.[41] Seventy percent might be a triumph for the coders who developed the program, but when placed into the context of the justice system, that inaccuracy rate of thirty percent becomes concerning.

Large language models do not seem to fare any better when it comes to accuracy. LLMs are deep learning algorithms, a sub-type of ML that uses an artificial neural network meant to simulate how the brain links disparate ideas.[42] They are designed to process natural language inputs, analyze the patterns and connections between words, and predict how certain sentences will end.[43] The model is trained through exposure to written language via books and articles, and as it analyzes the data, it absorbs grammar, facts, and sentence structure to the point that it can mimic human expression.[44] But this mimicry is not perfect. The LLM can learn biases present in the data it is exposed to, hallucinate information, and its reliability is often

[38] *See Traditional Programming vs Machine Learning*, INSIGHTSOFTWARE (Feb. 15, 2023), https://insightsoftware.com/blog/machine-learning-vs-traditional-programming [https://perma.cc/S3RQ-UN75].

[39] *Id.*

[40] *Id.*; *see also* Sara Brown, *Machine Learning, Explained*, MIT SLOAN SCH. OF MGMT. (Apr. 21, 2021), https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained [https://perma.cc/7USN-BAPA] (explaining that "programmers choose a machine learning model to use, supply the data, and let the computer model train itself to find patterns or make predictions," but the programmer still maintains some control over the process as they "can also tweak the model, including changing its parameters" to produce more accurate results).

[41] Kirsten Barkved, *How to Know if Your Machine Learning Model Has Good Performance*, ZAMS, https://www.zams.com/blog/machine-learning-model-performance [https://perma.cc/VK9Q-NNKV] (last visited Feb. 18, 2025).

[42] *See What Is Machine Learning?*, *supra* note 35.

[43] Mike Priest, *Large Language Models Explained*, BOOST.AI, https://boost.ai/blog/llms-large-language-models/ [https://perma.cc/45A2-BPWX] (Feb. 20, 2024).

[44] *Id.*

called into question because "due to the innate[ly] unpredictable nature of these models, achieving absolute . . . accuracy is presently unattainable."[45] In theory, ML should catch and fix mistakes, but "current LLMs struggle to self-correct their reasoning," and "expecting these models to inherently recognize and rectify their reasoning mistakes is overly optimistic."[46]

Computers may have seen a rapid evolution over the decades, but this advancement is not indicative of scientists' ultimate triumph over the ones and zeroes nor of the creation of true "artificial intelligence." Several programmers have gone so far as to claim that their AI can successfully pass the Turing Test, but those claims have been widely challenged, in part due to inconsistencies in the test's administration.[47] While "[c]urrent AI systems excel in narrow domains," they "lack the ability to transfer knowledge and skills across different areas of expertise, a hallmark of human intelligence."[48] The stark limitations of AI are only increasingly coming to the forefront as even the largest technology companies have admitted an overall lack of meaningful progress in creating a truly intelligent system.[49] A research team at Apple concluded that "current AI models are 'not capable of genuine logical reasoning,'" and the problem is not one that lends itself to an easy solution, assuming there is any solution at all.[50] The team warns that these concerns should give people caution "as more and more trust is given to AI's 'intelligence,'" which often "isn't what it might appear."[51]

[45] *Id.*

[46] JIE HUANG ET AL., LARGE LANGUAGE MODELS CANNOT SELF-CORRECT REASONING YET 8 (2024).

[47] Sanksshep Mahendra, *Has Any AI Passed the Turing Test?*, A.I. PLUS, https://www.aiplusinfo.com/blog/has-any-ai-passed-the-turing-test/ [https://perma.cc/NBA2-9U63] (Jan. 30, 2024, 8:31 PM). Though it remains the most widely used benchmark in evaluating computer intelligence, many criticize the Turing Test's lack of standardized rules, the fact that it does not encompass all forms of human intelligence, and that it could be "passed by unintelligent machines that use tricks and deception to fool humans." *Id.*

[48] *Id.*

[49] *See* Matthias Bastian, *Truly Intelligent AI: Three Things Google's AI Chief Says Are Missing*, THE DECODER (Jan. 13, 2022), https://the-decoder.com/truly-intelligent-ai-three-things-googles-ai-chief-says-are-missing/ [https://perma.cc/XB52-ECJ9]; Ryan Christoffel, *Apple Researchers Ran an AI Test that Exposed a Fundamental 'Intelligence' Flaw*, 9 to 5 MAC (Nov. 1, 2024, 7:42 AM), https://9to5mac.com/2024/11/01/apple-researchers-ran-an-ai-test-that-exposed-a-fundamental-intelligence-flaw/ [https://perma.cc/58JA-NZ3N].

[50] Christoffel, *supra* note 49.

[51] *Id.* (describing a test in which AI could not solve a simple math problem when written in word form rather than with pure numbers and when clearly irrelevant information was included in the problem); *see also* Tim Hardwick, *AI Companies*

Concerns of AI takeover create an inflated and overstated idea of AI's true capabilities, painting a picture of computers that can think just as well, or even better, than humans. In reality, computers remain incapable of independent thought, of doing anything beyond following the instructions of a programmer, and AI software is far from being as objective and infallible as people might believe. The multitudes of documented errors in both functionality and AI's ability to draw conclusions is a startling prospect when such systems are used as definitive asserters of truth in court.

B. The History of Algorithmic and AI-Based Evidence in Court

1. Risk Assessment Algorithms and Machine Learning Evidence

Machine learning takes many forms in the court system, from risk assessment and facial recognition to fingerprint analysis and generative AI. When offered as evidence, its output is generally found to be admissible under the Federal Rules of Evidence (FRE), though that admissibility is sometimes dependent upon how the algorithm was created.[52]

While predictive algorithms have become commonplace in both professional and private settings, particularly recognizable in targeted advertising, it is their usage in criminal justice which has garnered increasing controversy, especially when it concerns risk assessment.[53] The most prevalent application of this technology involves predicting recidivism rates—the likelihood that someone convicted of a crime will someday reoffend.[54] Of the various programs designed to make such predictions, COMPAS "has been used to assess over one million individuals in the criminal justice system since it was developed in 1998," and its Recidivism Risk Scale "has been in use since 2000."[55]

Reportedly Struggling to Improve Latest Models, MACRUMORS (Nov. 13, 2024, 5:30 AM), https://www.macrumors.com/2024/11/13/ai-companies-struggle-improve-llms/ [https://perma.cc/ZU68-B5UV] ("Leading artificial intelligence companies . . . are facing 'diminishing returns' . . . . Silicon Valley's belief that more computing power, data, and larger models will inevitably lead to better performance . . . could be based on false assumptions.").

[52] *See* Patrick W. Nutter, Comment, *Machine Learning Evidence: Admissibility and Weight*, 21 U. PA. J. CONST. L. 919, 932 (2019) ("Machine learning output is likely admissible . . . . [T]he exact manner in which the algorithm was created or the way it would be used at trial may, in some cases, render it inadmissible.").

[53] *See* Dressel & Farid, *supra* note 4, at 2–3.

[54] *Id.* at 3.

[55] *Id.*

To form a prediction, "COMPAS relies upon two types of data: (1) data gathered from an offenders' [sic] official record by a criminal justice professional, and (2) offenders' responses to questions that may be administered via either a paper and pencil survey or interview with a professional."[56] The individual's criminal history, employment status, age, gender, ethnicity, history of substance abuse, community ties, and level of education are all taken into account.[57] The proprietary software then uses this input data and provides an estimation of the offender's risk of violence, recidivism, and non-compliance to compute an overall "risk" score reported as either low, medium, or high as compared to other offenders.[58] For years, this information has been used during pretrial proceedings and by judges to help inform decisions on sentencing.[59]

In *State v. Loomis*, the defendant, who was involved in a drive-by shooting, received a COMPAS score indicating he presented a high risk of recidivism.[60] The State referenced this score during oral arguments as a factor that should be considered for sentencing, and the court evidently agreed.[61] "You're identified, through the COMPAS assessment, as an individual who is at high risk to the community," the judge stated.[62] "I'm ruling out probation because of the seriousness of the crime and because your history . . . and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend."[63]

Cases like *Loomis* are not rare. In *Santos v. Macauley*, the Sixth Circuit Court of Appeals reviewed the decision of a lower court which had considered the defendant's COMPAS risk assessment score as a factor in determining sentencing.[64] In

[56] JENNIFER L. SKEEM & JENNIFER ENO LOUDEN, ASSESSMENT OF EVIDENCE ON THE QUALITY OF THE CORRECTIONAL OFFENDER MANAGEMENT PROFILING FOR ALTERNATIVE SANCTIONS (COMPAS) 8 (2007), https://cpb-us-e2.wpmucdn.com/sites.uci.edu/dist/0/1149/files/2013/06/CDCR-Skeem-EnoLouden-COMPASeval-SECONDREVISION-final-Dec-28-07.pdf [https://perma.cc/B792-QJDR].

[57] *Justice Served? Discrimination in Algorithmic Risk Assessment*, RSCH. OUTREACH (Sept. 19, 2019), https://researchoutreach.org/articles/justice-served-discrimination-in-algorithmic-risk-assessment/ [https://perma.cc/HDQ4-WP9Z].

[58] SKEEM & ENO LOUDEN, *supra* note 56, at 8, 18.

[59] *See* Dressel & Farid, *supra* note 4, at 2–3.

[60] State v. Loomis, 881 N.W.2d 749, 754–55 (Wis. 2016).

[61] *See id.*

[62] *Id.* at 755.

[63] *Id.*

[64] *See* Santos v. Macauley, No. 21-1076, 2021 U.S. App. LEXIS 22888, at *2 (6th Cir. Aug. 2, 2021).

denying the defendant's petition, the court explained that reasonable jurists would agree the sentencing was appropriate and the use of COMPAS was not unconstitutional.[65] The Appellate Division of New York's Supreme Court reviewed a decision by the state's Parole Board after it denied a prisoner's request to be released, citing various factors which included the prisoner's COMPAS score.[66] The court affirmed the Board's denial of parole, referencing the prisoner's predicted likelihood to return to substance abuse and his inability to prove that the Board prejudiced his rights.[67] In another, more recent case, the Court of Appeals for the Second Circuit held a petitioner's claim that his COMPAS report contained two prejudicial falsehoods was without merit, going through each contested data point to show the algorithm's assessment was not erroneous.[68] While not every court states its support of the program quite as explicitly as *Loomis*, case law makes it clear that judges are more likely to take the risk score into consideration for sentencing and parole than to question its reliability.

Facial recognition is another commonly used metric for evaluating a defendant's innocence or guilt. Described by Congress as a biometric surveillance system, facial recognition software takes multiple factors into account, including age, cosmetics, whether the individual underwent plastic surgery, the individual's pose, and the potential effects of substance abuse.[69] There are multiple programs available, each using proprietary software to match a subject's face with a database of millions, if not billions, of images from public sources and driver's license photos or mugshots.[70] A 2016 study published by the Georgetown Law Center on Privacy and Technology found that "[o]ne in two American adults is in a law enforcement face recognition network."[71] A match using facial recognition software is useful for police and criminal investigators, but the results themselves

[65] *See id.* at *4.

[66] Cassidy v. N.Y. State Bd. of Parole, 35 N.Y.S.3d 132, 134 (App. Div. 2016).

[67] *Id.*

[68] Amaker v. Schiraldi, 812 F. App'x 21, 24–25 (2d Cir. 2020).

[69] Karissa Key, *Pros and Cons of Facial Recognition Used in Criminal Cases*, PUMPHREY L. (Sept. 4, 2023), https://www.pumphreylawfirm.com/blog/pros-and-cons-of-facial-recognition-used-in-criminal-cases/ [https://perma.cc/6R7W-J5AP].

[70] *Id.*

[71] Clare Garvie, Alvaro Bedoya & Jonathan Frankle, *The Perpetual Line-Up: Unregulated Police Face Recognition in America*, GEO. L. CTR. PRIV. & TECH. (Oct. 18, 2018), https://www.perpetuallineup.org [https://perma.cc/YNA5-BDBZ].

are not generally admissible as concrete evidence in court.[72] However, trial testimony regarding the use of facial recognition is permitted, and requests by defendants for discovery into the software that was used to help identify them are often denied.[73]

Finally, the rapid development of generative AI programs (GenAI), which are built upon text-generative LLMs,[74] means the possibility of litigation surrounding GenAI has become inevitable, automatically raising questions regarding the admissibility of such evidence.[75] Currently, GenAI has been involved in lawsuits involving privacy, tort, trademark, right of publicity, copyright, and facial recognition.[76] While criminal cases involving GenAI are not currently among these litigation trends, criminal justice is certainly not immune to the concerns of growing GenAI or the possibility of pleadings being written based on computer input.[77] Furthermore, the underlying concerns of ML software producing inaccurate results apply to each form of algorithmic-driven evidence and must be addressed so judges can make better-informed decisions in the future.

2. Probabilistic Genotyping

Probabilistic genotyping has emerged as a new tool employed by prosecutors, rising to prominence in part due to the "CSI Effect," which refers to a jury's greater tendency to find defendants guilty when DNA evidence is produced that ties them to the alleged crime.[78] Because probabilistic genotyping can supposedly isolate a single suspect's DNA among a sample containing DNA from multiple individuals, it can be used as

[72] *See* People v. Reyes, 133 N.Y.S.3d 433, 436–37 (Sup. Ct. 2020) ("[A] facial recognition 'match' has never been admitted at a New York criminal trial as evidence that an unknown person in one photo is the known person in another.").

[73] *See id.* at 435.

[74] Elizabeth Bell, *Generative AI vs. Large Language Models (LLMs): What's the Difference?*, APPIAN (Sept. 19, 2024), https://appian.com/blog/acp/process-automation/generative-ai-vs-large-language-models.html [https://perma.cc/V2HN-C2VP].

[75] *See* Maura R. Grossman et al., *The GPTJudge: Justice in a Generative AI World*, 23 DUKE L. & TECH. REV. 1, 1–2, 4 (2023).

[76] Christopher J. Valente et al., *Recent Trends in Generative Artificial Intelligence Litigation in the United States*, K&L GATES (Sept. 5, 2023), https://www.klgates.com/Recent-Trends-in-Generative-Artificial-Intelligence-Litigation-in-the-United-States-9-5-2023 [https://perma.cc/7YSR-DATN].

[77] *Id.*

[78] *See* Daniel P. Mooney, *The Rise of Probabilistic Genotyping Causing the Fall of DNA Evidence*, MSBA (Sept. 21, 2022), https://www.msba.org/site/site/content/News-and-Publications/News/General-News/The-Rise-of-Probabilistic-Genotyping-Causing-the-Fall-of-DNA-Evidence.aspx [https://perma.cc/ZK8R-6RPN].

evidence in a greater number of cases where uncontaminated samples are unavailable.[79] However, this new approach to genotyping differs greatly from previous methods of DNA analysis.

Traditional forensic DNA analysis typically involves either loci (physical locations on a chromosome) that contain Variable Numbers of Tandem Repeats (VNTRs), or polymerase chain reaction (PCR) based analyses.[80] VNTRs are regions of DNA with large quantities of alleles—alternative versions of a gene—and, as a result, are "particularly convenient as markers for human identification" because of the high level of variation between any two individuals.[81] By extracting DNA from a sample, running it through an electrified gel assay, and comparing the fragment lengths of VNTRs, scientists can determine whether two DNA samples are a match.[82] If there is not enough of an initial sample for this method, PCR is used to "greatly amplify[] a short segment of DNA" so that alleles can be identified and compared.[83] "PCR-based methods permit the analysis of extremely tiny amounts of DNA," but that sample must be from a single individual to prevent the risk of contamination and false results.[84] These traditional methods of DNA fingerprinting are highly reliable when analyzing evidence containing a DNA sample from one person, but it becomes far more complicated when mixed samples are at issue.[85]

Technological improvements have reduced the sample size needed for analysis, but because of this heightened sensitivity, DNA from multiple individuals is often detected.[86] Ordinary DNA analysis would be unable to accurately separate each genetic profile, but probabilistic genotyping "takes incomplete or otherwise inscrutable DNA left behind at a crime scene, often in minuscule amounts, and runs it through a software program that calculates how likely it is to have come from a particular

[79] *See id.*

[80] *See* NAT'L RES. COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 1, 4, 21, 216 (1996).

[81] *Id.* at 14–15.

[82] *See id.* at 15–17.

[83] *Id.* at 21.

[84] *Id.* at 23.

[85] *See NIST Publishes Review of DNA Mixture Interpretation Methods*, NIST (June 9, 2021), https://www.nist.gov/news-events/news/2021/06/nist-publishes-review-dna-mixture-interpretation-methods [https://perma.cc/G26V-MZAQ].

[86] JOHN M. BUTLER ET AL., DNA MIXTURE INTERPRETATION: A NIST SCIENTIFIC FOUNDATION REVIEW 11–12 (2024), https://nvlpubs.nist.gov/nistpubs/ir/2024/NIST.IR.8351.pdf [https://perma.cc/GE6M-WZLZ].

person."[87] This calculation is not perfect. According to the National Institute of Standards and Technology, while laboratories generally come to the same result when analyzing "high-quality, single-source samples," multiple interlaboratory studies over the last twenty years reveal "a wide range of results when interpreting the same *DNA mixtures*."[88]

At the forefront of this new genotyping technology is TrueAllele—a proprietary software created and sold by Cybergenetics.[89] Starting with a mixed DNA sample, the software "propose[s] tens of thousands of possible individual DNA profiles . . . [and provides] a 'likelihood ratio' that expresses the chance the suspect's DNA is in the evidence sample, relative to a random person in the population."[90] The likelihood ratio is not a conclusive match, but it provides compelling evidence and has been used in criminal trials since 2009[91] with forty crime labs approving it for use without reviewing its source code.[92] As

[87] Lauren Kirchner, *Powerful DNA Software Used in Hundreds of Criminal Cases Faces New Scrutiny*, THE MARKUP, https://themarkup.org/news/2021/03/09/powerful-dna-software-used-in-hundreds-of-criminal-cases-faces-new-scrutiny [https://perma.cc/D64H-A4TS] (Mar. 9, 2021, 9:59 AM).

[88] BUTLER ET AL., *supra* note 86, at 12.

> Distinguishing one person's DNA from another's in these mixtures, estimating how many individuals contributed to the recovered DNA sample, not knowing whether the DNA is associated with a crime or is from contamination, or whether the findings support the presence of a trace amount of suspect or victim DNA make DNA mixtures inherently more challenging to interpret than single-source samples. These issues, if not properly considered and communicated, can lead to misunderstanding the strength and relevance of the DNA evidence in a case.

Id. at 21.

[89] *See* Jouvenal, *supra* note 3.

[90] *Id.*

[91] In 2009, TrueAllele evidence was admitted for the first time, leading to a conviction of first-degree murder for the defendant who then appealed on the basis that testimony regarding TrueAllele should have been excluded because:

> (1) "as of the date of the pre-trial hearing, no forensic laboratory in the United States used Perlin's TrueAllel [sic] method in analyzing a mixed sample of DNA for forensic purposes"; (2) "the TrueAllel [sic] system had never been used in a court of law in any jurisdiction in the United States on a mixed DNA sample to give a likelihood ratio"; and (3) no outside scientist can replicate or validate Dr. Perlin's methodology because his computer software is proprietary.

Commonwealth v. Foley, 38 A.3d 882, 888–89 (Pa. Super. Ct. 2012) (alteration in original) (citation omitted). Affirming the admission of TrueAllele evidence, the court found there was no "legitimate dispute regarding the reliability" of the evidence, novelty of a scientific method is not based on its prior usage in court, and "scientists can validate the reliability of a computerized process even if the 'source code' underlying that process is not available to the public." *Id.* at 889.

[92] Jouvenal, *supra* note 3.

of this year, judges have ruled TrueAllele evidence to be admissible in over fifty cases after it was challenged by defendants at both the state and federal level.[93] As TrueAllele evidence increasingly arises in court, judges need to be prepared to make just and fair determinations about its reliability, which is only possible through an in-depth examination of the software.

III. PROBLEMS

A. Lost Transparency: Black Box Algorithms and Biased Reports

It is said that justice is blind, but in a very real sense, judges should not be. When a person's liberty hinges on evidence produced by a computer, understanding the process between the initial input data and its final output is vital to protecting justice. Yet this process is often shrouded in secrecy and hidden behind impenetrable walls in an all-too-common practice known as "black box" algorithms.[94] A program's source code dictates each action the program takes, revealing how and why it came to a certain conclusion. A lack of transparency in this regard is equivalent to an expert witness asking the court to simply take their word on something without any further explanation. And when validation studies of a program's effectiveness are offered in lieu of this explanation, they are often tainted by implicit bias. As a result, attorneys, criminal defendants, and even judges are starting to become more vocal in their protests against black box evidence, and systems like TrueAllele and COMPAS are leading offenders.[95]

These protests have arisen out of the backdrop of admissibility rules that currently govern scientific and computer-driven evidence—rules which some argue are ineffective at regulating AI. Generally, when scientific evidence is offered in trial, it is accompanied by an expert witness to explain their findings to the jury. The judge's decision whether to admit such evidence is guided by the FRE and standards explicated in two seminal cases: *Daubert v. Merrell Dow*

[93] *See* *TrueAllele Admissibility*, CYBERGENETICS, https://www.cybgen.com/information/admissibility/page.shtml [https://perma.cc/U6PT-C9WV] (last visited Apr. 9, 2025).

[94] *See* Christina Swarns, *When Artificial Intelligence Gets It Wrong*, INNOCENCE PROJECT (Sept. 19, 2023), https://innocenceproject.org/when-artificial-intelligence-gets-it-wrong/ [https://perma.cc/XNQ7-ZHBG].

[95] *See* Jouvenal, *supra* note 3.

Pharmaceuticals, Inc. and *Frye v. United States*.[96] Decided in 1923, *Frye* is a District of Columbia Circuit case in which the court held that admissibility is guided by whether the principle or method from which "the deduction is made . . . [is] established to have gained general acceptance" in the relevant scientific community.[97] Seventy years later, this rule was superseded by the Supreme Court's decision in *Daubert* which shifted general acceptance in the scientific community from the only test to one of several factors judges should consider, including whether the technique can be tested for reliability, its error rate, and whether it was subject to peer review.[98] Currently, federal courts exclusively follow the *Daubert* rule while state courts remain split between the two.[99] Under either standard, black box evidence is not per se invalid.[100] While these rules may seem to be a logical foundation of admissibility in a bubble, when applied to real cases involving defendants being convicted by algorithms, cracks begin to emerge.

After a 2014 robbery at a Virginia gas station, investigators were stumped for a lead when DNA analysis on the victim's shirt resulted in zero hits.[101] Four years later, the shirt was re-tested, and improved analytical techniques revealed residual DNA of at least three individuals, but only one-third of the necessary genetic markers were present to determine a match with any one person.[102] Under traditional analysis, the evidence was a dead end; using TrueAllele, a defendant was identified and charged. There was no other evidence directly linking the defendant to the crime, and he always maintained his innocence, claiming he had never stepped foot in the county where the robbery was

[96] Anjelica Cappellino, Daubert *vs.* Frye*: Navigating the Standards of Admissibility for Expert Testimony*, EXPERT INST., https://www.expertinstitute.com-resources/insights/daubert-vs-frye-navigating-the-standards-of-admissibility-for-expert-testimony/ [https://perma.cc/Z9TM-ZPCE] (Apr. 11, 2022).

[97] Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923).

[98] Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 592–94 (1993). Specifically, *Daubert* instructs that when "[f]aced with a proffer of expert scientific testimony, . . . the trial judge must . . . . [make] a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid." *Id.* at 592–93. Pertinent considerations include: (1) whether the relevant theory or technique can be (and has been) tested; (2) whether it has been subjected to peer review and publication; (3) the theory or technique's known or potential error rate; (4) the existence and maintenance of standards controlling its operation; and (5) whether the theory or technique has attracted widespread acceptance within the relevant scientific community. *Id.* at 593–94.

[99] Cappellino, *supra* note 96.

[100] *See* Nutter, *supra* note 52, at 949.

[101] *See* Jouvenal, *supra* note 3.

[102] *See id.*

committed.[103] The public defender assigned to his case argued "it would be impossible to assess whether TrueAllele had correctly identified [the defendant] . . . without the program's source code."[104] When faced with reports questioning the program's reliability, the attorney went on to say: "We shouldn't be using the criminal justice system as a proving ground for new technologies, especially when the makers of these technologies are keeping how they work secret."[105] The defendant was indicted for robbery and use of a firearm in 2019, and in 2021, the Virginia Court of Appeals reversed the grant of bail, concluding the lower court had erred in finding he was not a danger to the community.[106] However, concerns over the use of TrueAllele were not addressed.[107] His attorney was not the first nor the last to protest the use of TrueAllele and the pervasive secrecy surrounding the software.

In *People v. Wakefield,* the defendant was convicted of first-degree murder and first-degree robbery after his DNA was identified on several items taken from the crime scene.[108] Of the evidence analyzed, lab technicians discovered a complex mixture of multiple DNA samples from which the "defendant could not be excluded," but they were ultimately unable to determine whose DNA was present with any certainty.[109] Then the data was sent to Cybergenetics.[110]

> TrueAllele concluded that it was 5.88 billion times more probable that defendant was a contributor to the mixture on the amplifier cord than an unrelated black person, . . . 170 quintillion times more probable . . . [regarding] the outside rear shirt collar, . . . and 303 billion times more probable . . . [regarding] the mixture on the outside front shirt collar.[111]

With those numbers being presented as definitive, a jury would likely have no trouble finding the defendant guilty.

Before *Wakefield* went to trial, a *Frye* hearing was held to determine admissibility of the TrueAllele evidence, and defense

[103] *See id.*

[104] *Id.*

[105] *Id.*

[106] Commonwealth v. Watson, No. 1284-20-4, 2021 WL 2324262, at *1–2, *5–7 (Va. Ct. App. June 8, 2021).

[107] *See id.* at *2, *5–7.

[108] People v. Wakefield, 195 N.E.3d 19, 21–22, 26 (N.Y. 2022).

[109] *Id.* at 21–22.

[110] *Id.* at 22.

[111] *Id.*

counsel cited statements from Dr. Ranajit Chakraborty, a scientist specializing in evaluating methods of forensic DNA analysis.[112] Dr. Chakraborty explained TrueAllele is a "novel innovation" that has not gained "general acceptance in the scientific community," and although the program was approved by the New York State Commission on Forensic Science DNA Subcommittee, the committee had not been given proof of its analysis of complex DNA mixtures.[113] Overall, the program "ha[d] not been adequately validated for the type of casework [to which] it [was then] being applied, [and] . . . in the absence of disclosure of the source code . . . and the underlying assumptions programmed into the system, 'TrueAllele cannot be meaningfully validated.'"[114] In return, the People called several witnesses to advocate for TrueAllele's reliability, including its creator Dr. Mark Perlin, who argued that the software has been the subject of many peer-reviewed papers and underwent twenty-five validation studies.[115] Specifically calling out the software's black box nature, the defense cross-examined those witnesses, revealing that "laboratory analysts lack a complete understanding of how the . . . system works . . . [and therefore] would be [un]able to testify in court."[116] The court ultimately found the evidence to be admissible, determining TrueAllele is generally accepted in the scientific community.[117]

The studies referenced by Dr. Perlin in *Wakefield* are available on Cybergenetics' website, and a quick glance reveals one striking similarity between them—Dr. Perlin was a co-author in twenty-one out of the twenty-three listed journal

[112] *Id.* at 23; *see also* Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923) (holding that the test for admissibility of scientific evidence is whether the method at issue has "gained general acceptance in the particular field in which it belongs").

[113] *Wakefield*, 195 N.E.3d at 23.

[114] *Id.* (first and second alterations in original). A 2024 National Institutes of Standards and Technology study had similar concerns, explaining that while there have been validation studies on probabilistic genotyping since 2014, the information found in these publications is often lacking "specific details about the samples, including the assigned [likelihood ratio (LR)] values . . . [such that] reasons for differences [among mixed DNA sample analyses] cannot be independently assessed." BUTLER ET AL., *supra* note 86, at 82, 90. A further problem is the fact that "reliability" is a more subjective than objective inquiry when it comes to LRs, and "an assessment of reliability . . . for global forensic cases [is] not feasible for LR values assigned by [probabilistic genotyping] systems . . . in large part because there is no true LR." *Id.* at 15.

[115] *Wakefield*, 195 N.E.3d at 23–24.

[116] *Id.* at 25.

[117] *Id.*

publications.[118] The hallmark of a reliable peer-reviewed study is that it offers an independent, unbiased assessment of a particular scientific process, ensuring data was not altered or made in error by the original authors to create an artificially better result.[119] Further compounding the issue is the fact no laboratory assessing TrueAllele was given access to the program's source code, and scientists involved in determining its accuracy have admitted their lack of understanding as to how the program functions.[120] If the algorithms at issue were hidden from the very experts on which courts rely to provide an opinion, then any conclusions pertaining to the software's acceptance cannot be trusted as wholly accurate.

When arguments about general acceptance in the scientific community and due process fail, defense teams have attempted to exclude black box evidence on the grounds it violates a defendant's Sixth Amendment right to confront the witnesses against him—otherwise known as the Confrontation Clause.[121] After a 2013 double homicide in Pennsylvania, a bandana discovered at the crime scene was sent to Cybergenetics to determine if the defendant's DNA was among the mixed sample

[118] *Publications*, CYBERGENETICS, https://www.cybgen.com/information/admissibility/page.shtml [https://perma.cc/UAP2-4WXM] (last visited Apr. 9, 2025).

[119] *See Responsibilities in the Submission and Peer-Review Process,* INT'L COMM. OF MED. J. EDS., https://www.icmje.org/recommendations/browse/roles-and-responsibilities-responsibilities-in-the-submission-and-peer-peview-process.html [https://perma.cc/3SY9-2CDJ] (last visited May 4, 2025).

[120] *See* Jouvenal, *supra* note 3.

> [A] judge at a previous trial . . . asked a scientist trained on TrueAllele if she could independently reproduce the results of the program . . . [and she] replied: "It would take me years to try, and I don't know that I could do it." [The scientist] went on to testify that she wouldn't be able to detect low-level errors in TrueAllele's analysis either.

Id.

[121] *See, e.g.*, *Wakefield*, 195 N.E.3d at 26.

> Defendant . . . assert[ed] that the TrueAllele Casework System was the witness and that he needed the source code to effectively cross-examine that witness. . . . The court denied the request, stating that the issue defense counsel raised was a discovery issue and that defendant's ability to cross-examine Dr. Perlin . . . satisfied his right to confrontation.

Id.; Commonwealth v. Knight, No. 379, 2017 WL 5951725, at *6 (Pa. Super. Ct. Nov. 29, 2017) (finding the trial court properly denied discovery of TrueAllele's source code because it was not material to determining the program's reliability and defendant's right to confrontation was satisfied via cross-examination of Dr. Perlin); People v. H.K., 130 N.Y.S.3d 890, 897 (N.Y. Crim. Ct. 2020) (distinguishing STRMix from TrueAllele because Dr. Perlin describes TrueAllele as an "expert system" with "a certain degree of artificial intelligence," and therefore any Confrontation Clause concerns over STRMix were met by cross-examining the DNA analyst in a way that may not be satisfied with TrueAllele).

present on the cloth.[122] TrueAllele reported it was 5.7 billion times more likely the DNA belonged to the defendant than any other person, and the evidence was admitted at trial.[123] Arguing the evidence should be inadmissible without access to the source code, defense attorney Ken Haber intimated the Confrontation Clause, stating: "You can't cross-examine a computer. The Constitution demands, and justice requires, we be permitted to find out what the computer is doing to come up with its answer."[124] Haber's co-counsel, Noah Geary, had similar complaints, arguing Cybergenetics' refusal to release the source code was an "anathema to due process of law."[125] But when the defense submitted a discovery request for TrueAllele's source code, it was denied by the Court, which determined that the code is the "intellectual property of Cybergenetics," that it was neither material to the case nor necessary to evaluate the program's reliability, and that its release would cause irreparable harm to the company.[126]

TrueAllele does not stand alone in the field of secretive black box programs. Like Cybergenetics, Northpointe—the maker of COMPAS—considers the software's source code to be a trade secret and refuses to release the underlying algorithms for inspection.[127] Because the final recidivism score does not allow a user to understand how COMPAS reached that result, nor is a user able to confirm the program's supposed accuracy without its source code, "[t]he COMPAS system is not interpretable."[128] Defense teams have attempted to fight back against this secrecy in court, but without much success.

[122] *See Trials: Commonwealth of Pennsylvania v Michael Robinson*, CYBERGENETICS, https://www.cybgen.com/news/cases/Pennsylvania-v-Michael-Robinson.shtml [https://perma.cc/9434-Y5DP] (last visited Apr. 3, 2025).

[123] Paula Reed Ward, *Legal Question: How Do You Cross-Examine a Computer?*, PITT. POST-GAZETTE (Aug. 29, 2016), https://www.post-gazette.com/news/science/2016/08/29/Legal-question-how-do-you-cross-examine-a-computer/stories/201608280021 [https://perma.cc/E77Z-PT3J].

[124] *Id.*

[125] *Id.*

[126] Memorandum Order at 1–3, Commonwealth v. Robinson, No. CC 201307777 (Pa. Ct. C.P. Feb. 4, 2016).

[127] *See* Rick Jones, *The Siren Song of Objectivity: Risk Assessment Tools and Racial Disparity*, NACDL MEDIUM (July 26, 2018), https://nacdl.medium.com/from-the-president-the-siren-song-of-objectivity-risk-assessment-tools-and-racial-disparity-fa5ccb0698a5 [https://perma.cc/3823-82XQ].

[128] Brandon L. Garrett & Cynthia Rudin, *Interpretable Algorithm Forensics*, 120 PROCEEDINGS NAT'L ACAD. SCIS., Oct. 2, 2023, at 1, 6.

A notable example of this strategy is found in *Loomis* wherein the defendant argued the proprietary nature of COMPAS violated his "due process right to be sentenced based on accurate information" because the validity of the COMPAS score cannot be judged without disclosure of "how the risk scores are determined or how the factors are weighed."[129] Finding his arguments unpersuasive, the court pointed to Northpointe's COMPAS manual which gives a broad overview of how the program functions and explained Loomis could still challenge the resulting score, thereby protecting due process.[130] The court then cited various validation studies purporting to show that COMPAS "is a sufficiently accurate risk assessment tool."[131] Yet, like the issues plaguing TrueAllele's validation studies, a review of risk assessment instruments in the United States revealed that "[i]n most cases, validity had only been examined in one or two studies . . . and, frequently, those investigations were completed by the same people who developed the instrument."[132]

Due process and constitutional concerns are a common refrain among attorneys and outspoken scientists[133] but often seem to fall on deaf ears as judges give deference to the business-motivated arguments of companies. The tension between protecting a defendant's right to fair justice and protecting proprietary trade secrets is only becoming more complicated as forms of AI continue to enter the courtroom at an increasingly fast rate.

B. When Machines Get It Wrong: Subjectivity and Inaccurate Results

There is a tendency to view machines as infallible arbiters of truth, particularly because, in theory, they should lack the subjective viewpoints and biases that influence human

[129] State v. Loomis, 881 N.W.2d 749, 760–61 (Wis. 2016).

[130] *See id.*

[131] *Id.* at 762.

[132] SARAH L. DESMARAIS & JAY P. SINGH, RISK ASSESSMENT INSTRUMENTS VALIDATED AND IMPLEMENTED IN CORRECTIONAL SETTINGS IN THE UNITED STATES 2 (2013) https://csgjusticecenter.org/wp-content/uploads/2020/02/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf [https://perma.cc/35CF-7P7U].

[133] *See* Ward, *supra* note 123 ("Dan Krane, a professor of biological sciences . . . wrote that while validation studies are important, it is the source code that serves to implement the underlying concepts of the program. 'Human experts are expected to explain how they arrive at a conclusion .this [sic] same expectation can and should apply to a computer program . . . .'").

decision-making. But AI and ML software are not without flaws, and those flaws manifest as faulty and incorrect output. Even a one percent error rate in a program's source code can "correspond to tens of thousands of errors in a single program."[134] When that error rate is applied to data being used to convict, the implications become startling, especially when keeping source codes hidden could result in errors going undetected. As explained by defense attorney Noah Geary, "[s]omething may be scientifically reliable, but that does not mean it is without flaws . . . [which] may rise to the level of reasonable doubt."[135]

A clear example of those flaws arose surrounding one of TrueAllele's largest market rivals—STRmix.[136] As a probabilistic genotyping software, STRmix claims to solve the same complicated DNA puzzles as TrueAllele, but it has undergone harsh scrutiny pertaining to its use in the Australian justice system.[137] After STRmix released its source code for inspection, Queensland authorities reported that coding errors were discovered which "affected DNA likelihood ratios in 60 cases" and prompted the replacement of STRmix evidence in twenty-four cases.[138] Though the makers of STRmix stress that each identified miscode was minor and inconsequential,[139] what may seem minor to a business focused on sales differs greatly from what a jury may view as minor in determining someone's guilt or innocence. A similar situation occurred in a U.S. criminal case when DNA evidence extracted from under the victim's fingernail was sent to both TrueAllele and STRmix in hopes of identifying the perpetrator.[140] When each software came to a different

[134] Nutter, *supra* note 52, at 940.

[135] Ward, *supra* note 123.

[136] *See* Kwong, *supra* note 8, at 292.

[137] *Id.* at 292–93.

[138] David Murray, *Queensland Authorities Confirm 'Miscode' Affects DNA Evidence in Criminal Cases*, THE COURIER MAIL (Mar. 20, 2015, 10:00 PM), https://www.couriermail.com.au/news/queensland/queensland-authorities-confirm-miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3f1c59039efd1a2ef55af92b [https://perma.cc/C8X4-FV2E].

[139] *See Summary of Miscodes*, STRMIX (May 23, 2018, 9:00 AM), https://strmix.com/news/summary-of-miscodes/ [https://perma.cc/TUH6-9JR6].

[140] *See* Douglass Dowty, *Judge Tosses Key Cutting-Edge DNA Before Potsdam Trial in 12-Year-Old Boy's Murder*, SYRACUSE (Aug. 29, 2016, 6:27 PM), https://www.syracuse.com/crime/2016/08/judge_tosses_cutting-edge_dna-before_potsdam-trial_in_12-year-old_boys_murder.html [https://perma.cc/M7JZ-RC7V].

conclusion after analyzing the same DNA sample, the judge ruled the evidence was inadmissible.[141]

If each program was as accurate as their marketing materials proclaim, they should reach the same conclusions when analyzing the same DNA since their respective companies purport to use equivalent methods of probabilistic genotyping. Yet inconsistencies between results is not uncommon. A study conducted by the Department of Criminology, Law and Society at the University of California, Irvine (UCI) used TrueAllele and STRmix to analyze a mixed DNA sample that traditional methods of analysis could not match.[142] Each program produced startling different results. TrueAllele presented four values comparing the DNA to various racial groups with a likelihood of a match between the sample and the defendant ranging from 1.2 million to 6.07 million times less probable than a "coincidental match" to another person.[143] STRmix determined the sample was twenty-four times more likely to have originated from two unknown contributors other than the defendant.[144] While both concluded the defendant's DNA was likely not in the mixture, TrueAllele's likelihood ratio was "larger by five to six orders of magnitude."[145] One's guilt or innocence should not hinge on which program was used to analyze evidence.

The author of the UCI study attempted to account for the discrepancy, pointing to variations in how each program sets certain values for analysis, as well as "misleading" ways in which the results are presented.[146] Specifically calling out TrueAllele, he explained, "[b]ecause Cybergenetics is using the terms 'match' and 'coincidence' to convey a meaning that is very different from what most people think those terms mean, and because Cybergenetics fails to explain this departure . . . I believe that Cybergenetics' LR [likelihood ratio] statement is inappropriate and misleading."[147] Overall, the study highlighted significant

[141] Lauren Kirchner, *Where Traditional DNA Testing Fails, Algorithms Take Over*, PROPUBLICA (Nov. 4, 2016, 8:00 AM), https://www.propublica.org/article/where-traditional-dna-testing-fails-algorithms-take-over [https://perma.cc/N4Q7-FKQ4].

[142] William C. Thompson, *Uncertainty in Probabilistic Genotyping of Low Template DNA: A Case Study Comparing STRMix and TrueAllele*, 68 J. FORENSIC SCIS. 1049, 1051 (2023).

[143] *Id.* at 1053.

[144] *Id.*

[145] *Id.* at 1054.

[146] *Id.*

[147] *Id.* at 1059. "Misleading" seems to be an appropriate term for the LRs expressed in probabilistic genotyping software. While most lay readers and lawyers might assume it

problems in both how these programs are implemented and how they are interpreted by lawyers and jurors, warning that statistical models resting on "unrealistic assumptions" will produce false results.[148]

In a paper published by the same author only a few months later, Dr. Thompson responded to criticism of his study, specifically calling out Dr. Perlin for making misleading statements regarding TrueAllele's reliability and for insinuating there were errors in his conclusions.[149] Explaining that his original article "raised a number of additional concerns about [TrueAllele] that Dr. Perlin and his colleagues failed to address,"[150] Dr. Thompson questioned Dr. Perlin's claim that TrueAllele is a fully automatic, Bayesian-statistics-based[151] system with "no need for analytic thresholds."[152] Unable to agree with Dr. Perlin's assessment that TrueAllele would always produce reliable, trustworthy data, Dr. Thompson invoked a sentiment common among computer scientists: "garbage in-garbage out."[153] It is unclear "at what point[] the

represents the likelihood of a proposition being true—of the defendant's DNA being present in the mixed sample—that is not the case. Instead, it represents the "ratio of the probability of the findings given [hypothesis one] is true," namely that the defendant's DNA contributed to the sample, "versus the probability of the findings given [hypothesis two] is true," that someone other than the defendant contributed to the sample. BUTLER ET AL., *supra* note 86, at 49–50. If the LR is one hundred, then the chance of seeing the exact DNA results present in the sample is one hundred times more likely to occur if the defendant's DNA is part of the sample rather than another person's. It does not mean that it is one hundred times more likely that the defendant's DNA is actually in the sample. The misinterpretation of this value is known as "transposing the conditional," or the "prosecutor's fallacy," in which a person "confuses 'the probability of the evidence given the propositions' with 'the probability of the propositions given the evidence.'" *Id.* at 50–51. For example, instead of DNA evidence, the evidence at issue is Earth's sky. And instead of the proposition being that the defendant's DNA is in the DNA evidence, your proposition is that the sky appears blue. The LR would represent the probability that you're looking at Earth's sky rather than the sky on Mars, given the proposition that the sky is blue. It does not tell you the likelihood of the sky actually being blue.

[148] *See* Thompson, *supra* note 142, at 1050, 1060.

[149] *See* William C. Thompson, *Author's Response*, 69 J. FORENSIC SCIS. 1519, 1519 (2024).

[150] *Id.* at 1521.

[151] Based on Bayes' theorem, which "is a mathematical formula that determines the conditional probability of any given event," Bayesian statistics is an approach to data analysis whereby "available knowledge regarding parameters in statistical models is updated using the information gathered from observed data." John Terra, *What Is Bayesian Statistics, and How Does It Differ from Classical Methods?*, CALTECH, https://pg-p.ctme.caltech.edu/blog/data-science/what-is-bayesian-statistics [https://perma.cc/5XKB-XT7C] (Aug. 14, 2024). This means that probabilities are continuously refined as more evidence becomes available. *See id.*

[152] Thompson, *supra* note 149.

[153] *Id.*

LRs produced by [TrueAllele] become garbage," and research has already suggested TrueAllele is *not* reliable for all types of DNA mixed samples.[154]

The larger problem with probabilistic genotyping like TrueAllele is the inherent subjectivity involved in certain aspects of the process—subjectivity that is now governed by computer software hidden behind claims of trade secrets. "Numerical results obtained from assigning LR values are dependent on the evidence available, statistical models applied, propositions selected based on case information, and the scientist making various judgments. . . . [which means] results vary based on [the] amount of information available and assumptions made."[155] The way in which the program is implemented directly affects the program's output and the corresponding result offered in court. In general, probabilistic genotyping systems compute LR based on:

> (1) *modeling choices* made by the system architect(s), (2) *data input choices* made by the analyst regarding an analytical threshold for calling peaks as alleles, selecting the number of contributors to the mixture for use in PGS calculations, and sometimes categorizing artifacts (e.g., pull-up peaks), (3) *proposition choices and assumptions* made by the analyst (e.g., use of unrelated individuals versus relatives, conditioning on a victim when analyzing an intimate sample, . . . and underestimating or overestimating the number of contributors), and (4) *population database choices* used by the laboratory to provide allele and genotype frequency estimates.[156]

When a forensic scientist makes those necessary choices and judgments, he or she can explain to the court which choices were made and why. When a computer running black box software does so, the court is left in the dark. In fact, the need for certain judgments to be made during the calculation process is a contributing factor as to why various probabilistic genotyping programs will come to vastly different conclusions when analyzing the same evidence.[157] And if there are flaws in the coding that

[154] *Id.* Studies have shown direct evidence of TrueAllele producing "falsely exculpatory" LRs, finding that a defendant's DNA was more likely to not be present in a mixed sample when, in reality, it *was* included in the sample. Validation studies cited by Dr. Perlin in response to these concerns do not address this particular issue and in fact "establish the opposite"—"[t]hey show that exculpatory results of this type are often NOT accurate and hence cannot be trusted." *Id.* at 1519–20.

[155] BUTLER ET AL., *supra* note 86, at 48.

[156] *Id.* at 51.

[157] *See id.* at 52–53.

directs the program on how to judge certain aspects of DNA evidence, there are corresponding flaws in the output—flaws that go unseen and unchecked under the current system for admitting machine evidence.

In the context of risk assessment tools, the potential for false conclusions has already been realized and proven in several studies analyzing the accuracy of their predictions. In 2016, *ProPublica* published a groundbreaking review of COMPAS, bringing the program under heavy inquiry after it was revealed to produce discriminatory and biased outputs.[158] Comparing two defendants, one African American with only juvenile misdemeanors and one Caucasian with armed robbery and attempted armed robbery convictions, COMPAS predicted the African American had a higher likelihood of recidivism.[159] Two years later, that defendant had no further charges while the Caucasian defendant was serving an eight-year prison sentence for robbery; COMPAS "got it exactly backward."[160]

Digging deeper into the program, *ProPublica* obtained the risk scores assigned to seven thousand defendants in Florida from 2013 to 2014 and found COMPAS had a success rate of only twenty percent when it came to accurately predicting which would go on to commit further crimes.[161] There were also "significant racial disparities" that remained unaffected by records of prior offenses, and "the scores ma[de] little sense even to defendants."[162] Taking *ProPublica*'s research a step further, Dr. Melissa Hamilton of the University of Surrey analyzed COMPAS's predictive validity in terms of gender, "proving that the tool overpredicts the risk for women to reoffend, therefore leading to unfair penalties for female offenders."[163] Dr. Hamilton

> Likelihood ratios are assigned and not measured. Different individuals may assign different LR values, even when using [probabilistic genotyping] systems, when presented with the same evidence because they base their judgments on different collection protocols, quantification systems, STR kit results, interpretation protocols, models, assumptions, or computational algorithms. For any given sample, there is no single, true likelihood ratio.

Id. at 54.

[158] *See* Angwin et al., *supra* note 6.

[159] *Id.*

[160] *Id.*

[161] *Id.*

[162] *Id.*; *see also* DESMARAIS & SINGH, *supra* note 132, at 49 ("[P]erformance within and between [risk assessment] instruments varie[s] considerably depending on the assessment sample, circumstances, and recidivism outcome.").

[163] *Justice Served? Discrimination in Algorithmic Risk Assessment*, *supra* note 57.

attributed this disparity in part to the fact the ML algorithm had been trained using samples of primarily male offenders, thereby limiting the scope of the program's capabilities and increasing the potential for bias.[164]

But it seems the more companies attempt to manipulate programming and data sets to increase objectivity and reduce bias, the less objective these programs become. Recently, Google launched an updated LLM called Gemini, which "produce[d] images of Black, Native American and Asian people when prompted – but refuse[d] to do the same for White people," citing racial concerns.[165] When Gemini was asked to show historical pictures of Nazi soldiers during World War II, it only produced images of people of color wearing the distinctive Nazi uniform, unable to understand the historical inaccuracy of the results.[166] Google apologized for the debacle and removed Gemini to "improve" its programming,[167] but the situation provides a clear, visual example of how coding can manifest unanticipated mistakes.

Empirical data shows there is not a single sector of ML or AI-driven software free from errors. While facial recognition has high accuracy in an ideal environment, real-world conditions are often far from perfect and result in error rates anywhere from

[164] *See id.*

[165] Nikolas Lanum, *Google Apologizes After New Gemini AI Refuses to Show Pictures, Achievements of White People*, FOX BUS. (Feb. 21, 2024, 2:32 PM), https://www.foxbusiness.com/media/google-apologizes-new-gemini-ai-refuses-show-pictures-achievements-white-people [https://perma.cc/E69B-UL5J].

[166] *See* Adi Robertson, *Google Apologizes for 'Missing the Mark' After Gemini Generated Racially Diverse Nazi*, THE VERGE (Feb. 21, 2024, 2:17 PM), https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical [https://perma.cc/QBV9-5JYN]. Programmers have attempted to fix inaccuracies such as this with self-correction tools, but "LLMs sometimes actually perform worse with self-correction measures," and "self-correction isn't consistently effective." Samantha Keefe & Thomas Gaitley, *AI Self-Correction*, LIONBRIDGE (Jan. 16, 2024, 9:30 AM), https://www.lionbridge.com/blog/translation-localization/ai-self-correction/ [https://perma.cc/3S3T-FKA2]. While certain issues associated with LLMs like Gemini are not wholly relevant to the operation of programs like TrueAllele or COMPAS, which do not use GenAI to produce output, their inaccuracies highlight a larger problem: how can any form of machine evidence be trusted when the very programmers designed to ensure accurate output are unable to fix even a problem as obvious as non-White Nazi officers? The building blocks of Gemini, TrueAllele, and COMPAS are the same—algorithms and lines of code telling the computer what to do—but just because errors are more clearly visible in GenAI does not mean errors do not exist in programs like COMPAS; simply, those errors are more easily hidden.

[167] *See* Robertson, *supra* note 166.

9.3% to 64%.[168] These programs have also been shown to misidentify female and minority populations at a disproportionate rate, often because of the skewed and incomplete information used to train the underlying algorithms.[169]

Gaining heightened media attention is the phenomenon of hallucinations plaguing the field of GenAI. While the nature of LLMs makes it difficult to determine exactly how often fabricated results are produced, estimates reveal hallucinations occur up to 27% of the time[170] and "leading AI experts aren't entirely sure what causes hallucinations."[171] In an extreme example of this issue, a New York attorney used a GenAI program to assist in writing his trial brief, and six of the referenced cases were entirely contrived "with bogus quotes and bogus internal citations."[172] When directly asked, the program assured the user that the cases were real and even provided full citations and instructions to locate them on Westlaw and LexisNexis.[173]

Hallucinations are not directly comparable to the types of problems plaguing machine evidence like TrueAllele, but the phenomenon raises an important point about the inherent trustworthiness, or lack thereof, of machine evidence, especially those with hidden source codes. It also raises an even larger issue inherent in any discussion about the use of AI in court—the implicit, misplaced trust lawyers (and judges) have in computer output. "If a computer said it, it must be true" seems to be the prevailing attitude, so much so that a lawyer relied on GenAI to provide him a trial brief without once considering that the

[168] *See* William Crumpler, *How Accurate Are Facial Recognition Systems – and Why Does It Matter?*, CTR. FOR STRATEGIC & INT'L STUD. (Apr. 14, 2020), https://www.csis.org/blogs/strategic-technologies-blog/how-accurate-are-facial-recognition-systems-and-why-does-it [https://perma.cc/W5X2-N7J4].

[169] *See* John McNichols, *How Do You Cross-Examine Siri if You Think She's Lying?*, AM. BAR ASS'N (May 24, 2022), https://www.americanbar.org/groups/litigation/resources/litigation-news/2022/how-do-you-cross-examine/ [https://perma.cc/N657-L28L].

[170] *See* Stefan Bardega, *Generative AI Hallucinations: How Often Do They Happen and Should Marketers Be Worried?*, IDX (Nov. 8, 2023), https://www.idx.inc/blog/technology/gen-ai-hallucinations [https://perma.cc/W7QV-KZFY].

[171] *Generative AI Hallucinations: Why They Occur and How to Prevent Them*, TELUS DIGIT. (July 6, 2023), https://www.telusinternational.com/insights/ai-data/article/generative-ai-hallucinations [https://perma.cc/8V2W-T7NS].

[172] Ramishah Maruf, *Lawyer Apologizes for Fake Court Citations from ChatGPT*, CNN BUS. (May 28, 2023, 3:28 PM), https://www.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html [https://perma.cc/5YZ4-VFZL].

[173] *See id.*

information in the brief could be inaccurate.[174] Trusting faulty citations is one thing; trusting faulty conclusions about a defendant's guilt or innocence is quite another.

Computer programming has greatly advanced over the decades, but it must be remembered that algorithms are in no way perfect. "The human is in the software in the source code,"[175] and humans make mistakes. Miscodes, incomplete training data, and underlying biases all contribute to incorrect output, making AI less "intelligent" than one might assume and highlighting the danger in blindly trusting a computer in court.

IV. Previously Proposed Solutions and Why None Are the Answer

A. Revising Current Admissibility Standards

When admitting AI-driven and machine evidence, it can be judged under one of three broad standards—"direct witness testimony, expert witness testimony, or measurement using established technology"[176]—but no method on its own provides the level of rigor necessary to ensure reliability. Most scholars seem to agree machine evidence should be evaluated in the same manner as previous technological advancements, but some have proposed revising current admissibility guidelines to carve out unique rules applicable to AI. Not only would this require amending the FRE or the *Daubert* and *Frye* standards, but it would not solve many of the overarching issues outlined in Part III.

Several authors have suggested revisions to either FRE 901, which governs authentication,[177] or rules 702 to 704, which provide guidelines for admitting expert testimony and implicate the judge-made rules in both *Daubert* and *Frye*.[178] As written,

[174] This sentiment only appears to be strengthening despite the multitude of studies demonstrating the dangers of relying on various forms of AI. Thomson Reuters, a company well-known for providing information and services within the legal profession, has published articles praising the advent of AI in the legal system, even going so far as to claim AI can "improve equity and reduce bias in judicial outcomes." Allyson Brunette, *Humanizing Justice: The Transformational Impact of AI in Courts, from Filing to Sentencing*, THOMSON REUTERS (Oct. 25, 2024), https://www.thomsonreuters.com/en-us/posts/ai-in-courts/humanizing-justice/ [https://perma.cc/UD9W-KJ92] ("Artificial intelligence (AI) tools are being introduced at every step of [the legal system] . . . . [to] improve[] efficiency and equity for defendants and their legal representation.").

[175] Ward, *supra* note 123.

[176] Paul W. Grimm, Maura R. Grossman & Gordon V. Cormack, *Artificial Intelligence as Evidence*, 19 NW. J. TECH. & INTELL. PROP. 9, 79 (2021).

[177] FED. R. EVID. 901.

[178] *See* FED. R. EVID. 702–04.

FRE 901 broadly governs how to authenticate evidence—how to prove a piece of evidence actually is what a party says it is.[179] Before the development of AI, this was a fairly simple task; now, when generated images and text are sometimes indistinguishable from the same materials produced by a human, it becomes far more complicated. FRE 702 to 704 lay out guidelines for the admissibility of expert testimony, including the expert's specialized knowledge in the subject and whether the testimony is based upon "sufficient facts or data" and "is the product of reliable principles and methods."[180] When scientific evidence was produced by a technician in a lab, this rule was a logical way of ensuring reliability. With AI, it is unclear whether any expert can truly have detailed knowledge as to how a black box program functions or testify regarding its reliability without access to its code, and an increasing number of scholars have begun to ask these pertinent questions.

Specifically addressing the problem of deepfakes, which are highly realistic AI-generated images, videos, and audio,[181] law professor Rebecca Delfino argues that the current FRE is incapable of meeting the challenges of this new category of evidence.[182] Like hallucinations, deepfakes are not a direct result of systems like TrueAllele or COMPAS, but they tend to show a clear, visual example of the pervasive problems posed by forms of AI—problems that are only exacerbated by programs purporting to definitively prove "who done it." Delfino critiques the current division of responsibility between the judge and jury for authenticating evidence, pointing to the fact jurors may be convinced by false AI evidence or allow personal skepticisms or biases to govern their decisions, and instead proposes tipping the balance in favor of judges.[183] FRE 901 "should be amended to add a new subdivision (c) . . . [that would] expand the gatekeeping function of the court by assigning the responsibility of deciding

[179] FED. R. EVID. 901(a) ("To satisfy the requirement of authenticating or identifying an item of evidence, the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is.").

[180] FED. R. EVID. 702(a)–(d).

[181] Ian Sample, *What Are Deepfakes – and How Can You Spot Them?*, THE GUARDIAN (Jan. 13, 2020, 5:00 PM), https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them [https://perma.cc/266Y-9VKT].

[182] *See* Rebecca A. Delfino, *Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery*, 74 HASTINGS L. J. 293, 332 (2023).

[183] *See id.* at 336–37, 341.

authenticity issues solely to the judge."[184] The author argues this expanded role of the judge's fact-finding responsibilities is not without precedent, citing traditional English law which gives judges broad plenary power "to decide all questions of fact conditioning the admissibility of testimony."[185] This new division of authority to determine authenticity is meant to solve the problem of admitting realistic deepfakes and the potentially cost prohibitive nature of proving whether an image is real or AI-generated.[186]

Although Delfino articulates several poignant issues regarding FRE 901's application to AI evidence, her solution is not without flaws. For Delfino's plan to be most effective, the judge would need a requisite understanding of AI and deepfake technology in order to properly determine whether such evidence was more likely authentic or falsified. Yet the language of the suggested amendment is silent on this issue, and the article overall offers no proposed mechanism for judges to obtain that expertise.[187] This problem is exacerbated by the fact that according to Rule 2.9(c) of the Model Code of Judicial Conduct—a rule over thirty states have adopted[188]—judges are not authorized to independently research facts.[189] They therefore could not gain a meaningful understanding of deepfake technology adequate enough to singlehandedly decide issues of AI admissibility without opinions from an expert.

Taking a broader approach to address AI generally, another law professor, Victor Metallo, focuses on the rules surrounding expert scientific testimony and suggests amending the *Daubert* standard as well as the FRE's guidelines for expert testimony

[184] *Id.* at 341.

[185] *Id.* at 342.

[186] *See id.* at 340.

[187] *See id.* at 341.

> Notwithstanding subdivision (a), to satisfy the requirement of authenticating or identifying an item of audiovisual evidence, the proponent must produce evidence that the item is what the proponent claims it is in accordance with subdivision (b). The court must decide any question about whether the evidence is admissible.

Id. (suggesting amendment to FED. R. EVID. 901(c)).

[188] KEITH R. FISHER, NAT'L CTR. FOR STATE CTS., NEW ABA ETHICS OPINION EXPLORES THE PROHIBITION ON INDEPENDENT FACT RESEARCH BY JUDGES 2 (2018), https://www.ncsc.org/__data/assets/pdf_file/0022/19309/new-aba-ethics-opinion-explores-prohibition-independent-fact-research-by-judges.pdf [https://perma.cc/G66H-LX3B].

[189] MODEL CODE OF JUD. CONDUCT r. 2.9(c) (AM. BAR ASS'N 2020).

under FRE 702.[190] Metallo argues the factor-based *Daubert* test, which was designed to ensure "that any and all scientific testimony or evidence admitted is not only relevant, but reliable,"[191] is only adequate for machine evidence akin to calculators that simply speed up a human function.[192] It is, however, insufficient to tackle situations "where the machine is the unique source of knowledge, [and] where the human being might be just a vector delivering that knowledge."[193] To solve this conundrum, Metallo proposes FRE 702 should be amended to include "reliability requirements for AI" while also ensuring AI cannot wholly replace human expert testimony, only assist it.[194]

However, this proposed amendment does not fully address the root problem concerning AI evidence and could potentially create confusion or inconsistency—an outcome contrary to Metallo's goal of avoiding conflicting admissibility decisions under the current FRE.[195] Implementing "reliability requirements" could be highly beneficial, but the phrase lacks explanation as to what those requirements might be or how the courts should weigh various factors that may point towards a program's reliability. Should there be a delineated error rate over which no evidence is admissible? Would certain types of ML be afforded more protection than others? Without answering these questions, there can be no assurance such a change will effectively result in more consistent decisions.

Metallo also calls for a further amendment to the FRE which would permit the court to deem testimony inadmissible under the following circumstances:

> (1) a judge cannot take judicial notice of an Al process; or (2) where a party has not proffered an engineer to assist in explaining the Al's processes to a jury; or (3) where the Al has reached a point that "black box" processes cannot be explained by human testimony, because Al has adapted the ability to program itself.[196]

Analyzing this secondary proposal, more questions begin to emerge. Basing discretion on whether the judge can "take judicial

[190] *See* Victor Nicholas A. Metallo, *The Impact of Artificial Intelligence on Forensic Accounting and Testimony—Congress Should Amend "The* Daubert *Rule" to Include a New Standard*, 69 EMORY L.J. ONLINE 2039, 2041–42 (2020).

[191] Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 589 (1993).

[192] Metallo, *supra* note 190, at 2048–49.

[193] *Id.*

[194] *Id.* at 2060.

[195] *See id.*

[196] *Id.* at 2061.

notice of an AI process" does not consider black box issues wherein no one can take notice of a program's process, nor does it address the judge's inability to research facts about such a process on their own. But arguably the most problematic aspect is subsection (3). As written, the amendment suggests AI may be capable of programming itself, yet most computer scientists agree that while certain AI models have been trained to mimic code written by humans, they cannot learn it to the degree necessary to create new programs.[197] "The computer appears to 'understand' things . . . [but] it wouldn't understand ANY of that if human programmers hadn't first painstakingly taught it how," one software developer explains.[198] "AI systems can't even *learn* on their own. . . . [and e]ven with new-fangled quantum hardware, I would seriously question a computer's ability to come up with innovative code . . . Behind every successful AI is a programmer rolling their eyes."[199]

The misunderstanding of the tenets of computer science implicated by this proposal demonstrates the danger of entrusting the evaluation of machine evidence solely to lawyers and judges whose areas of expertise are in a far different field. While the current rules may not address AI outright, no suggested revision appears to be without flaws that could further complicate an already convoluted area of evidence.

B. Maintaining the Status Quo

Rather than developing new admissibility rules each time there is a technological advancement that affects the legal system, others suggest keeping the rules as they are and simply adjusting their application when AI-driven evidence is at issue. The wheels of justice are known to move slowly; those of science try to break the speed limit. Because the FRE and its state counterparts are revised infrequently and via an extremely lengthy process, it is impractical to set a standard of revising the rules each time a new form of evidence is developed.[200] As a result, certain scholars argue there is nothing inherently inadequate about applying the rules as they stand. But leaving the status quo in place without developing a new method of

[197] *See* Tim Baker, *What Artificial Intelligence Can't Do*, MEDIUM (Sept. 12, 2023), https://medium.com/codex/what-artificial-intelligence-cant-do-b92b4ddcf8b3 [https://perma.cc/C2M9-UBF6].

[198] *Id.*

[199] *Id.*

[200] *See* Grossman et al., *supra* note 75, at 16.

evaluation better equipped to handle this emerging field of evidence is not without its problems as well.

The standards for expert testimony provide a clear example of this problem. The FRE allows for such testimony if the expert's "scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue,"[201] and if their opinion is based on "facts or data in the case that the expert has been made aware of or personally observed."[202] The Supreme Court has explained these rules grant expert witnesses "testimonial latitude . . . on the assumption that the expert's opinion will have a reliable basis in the knowledge and experience of his discipline."[203] While this assumption may be acceptable in other areas of expertise, it rests on shakier ground with programs that do not require user input, or even user understanding, to reach a result. In the Sixth Circuit case *U.S. v. Ganier*, the government argued against the admissibility of expert testimony based on computer search results, contending such testimony "is not based on scientific, technical, or other specialized knowledge, but is simply lay testimony available by 'running . . . software, obtaining results, and reciting them.'"[204] Using AI is no different than using a search engine in this respect—a user inputs information and receives an answer without necessarily knowing or understanding the steps in between. As written, the FRE does not address this shift from the traditional definition of expert testimony. Yet many scholars propose that the current system of deciding admissibility does not need alteration and is more than capable of handling AI.

Focusing on the admissibility of ML evidence, author Patrick Nutter surveys the existing rules before concluding "there is nothing inherently inadmissible about ML evidence under the Federal Rules of Evidence, the Fifth Amendment, or the Sixth Amendment."[205] He recognizes the potential dangers in allowing such evidence to have blanket admissibility, but rather than amending any existing standard, he suggests allowing the trier of fact to decide for themselves what weight to give computer evidence.[206] Briefly touching on the unexplainability of much of

[201] FED. R. EVID. 702(a).

[202] FED. R. EVID. 703.

[203] Kumho Tire Co. v. Carmichael, 526 U.S. 137, 138 (1999).

[204] U.S. v. Ganier, 468 F.3d 920, 925 (6th Cir. 2006).

[205] Nutter, *supra* note 52, at 949.

[206] *See id.* at 919.

that evidence, especially concerning black box programs, Nutter determines the "Sixth Amendment merely requires that the evidence be introduced with expert testimony."[207]

His paper provides an excellent outline of the significant evidentiary issues plaguing ML but does not seem to give an in-depth examination as to how the current setup is sufficiently designed to address them. Though the author cautions that judges and juries should be wary when weighing the strength of computer evidence, there is no explicit guidance regarding how they should do so nor whether certain types of AI require more detailed analysis than others. Overall, this position essentially leaves the situation exactly as it stands, solving none of the problems many scholars claim will soon be impacting most litigators.[208]

Taking a similar stance to Nutter, the authors of a paper, written in part by former U.S. District Court Judge Paul Grimm, argue the FRE is adequate for evaluating machine evidence without amendment "provided [the rules] are applied flexibly."[209] The authors emphasize the importance of demonstrating whether the AI evidence can be trusted as accurate before judges are then free to use their already broad discretion in deciding questions of admissibility.[210] Specifically, if a party plans to offer AI evidence, they must do so in advance of trial to ensure it meets "adequate thresholds of validity and reliability" before being presented to the jury.[211] When black box evidence is at issue, and the program's source code is not revealed, the authors explain that the party arguing for its admissibility must demonstrate validity and reliability some other way.[212]

Although reliability thresholds would be an important step forward, their paper leaves many questions unanswered and does not provide clear guidelines on how to evaluate certain types of machine evidence nor how a party would go about adequately proving a program's reliability. It places the burden entirely on

[207] *Id.* at 958.

[208] *See id.* at 919 ("Artificial intelligence ('AI') is gaining traction in legal practice. How prosecutors prioritize which crimes to prosecute, sift through mountains of documents, and establish reasonable suspicion can all reasonably be expected to change with coming AI technologies."); *see also* Delfino, *supra* note 182, at 297 (arguing that AI "will soon make trial attorneys' and judges' jobs significantly more challenging" and will require "additional measures" to evaluate).

[209] Grimm, Grossman & Cormack, *supra* note 176, at 85.

[210] *See id.* at 104.

[211] *Id.* at 89.

[212] *Id.*

the court system, yet the authors themselves admit, "When it comes to technical evidence like AI, the judge often is in a battle of wits unarmed," and leaving the obligation to attorneys "can be a challenge for lawyers who . . . [are] not specialists in the many scientific and technical disciplines that underlie AI systems."[213] They acknowledge the inherent difficulties in relying on current admissibility standards, yet ultimately argue those same standards are more than capable of handling ever-evolving AI evidence.

This position is also troublesome when it comes to excluding hearsay. FRE 801 defines hearsay as an assertion made outside the current trial, which is offered to "prove the truth of the matter asserted in the statement."[214] Because "statement" is defined as a "person's" assertion,[215] the authors distinguish AI-generated output because it is not a direct human assertion and therefore any issues of hearsay are irrelevant.[216] To support this position, the authors cite multiple federal cases that echo the sentiments of the Fourth Circuit, which determined "[o]nly a *person* may be a declarant and make a statement," and therefore, nothing "said" by a machine can qualify as hearsay.[217] As such, the authors dismiss the concept out of hand, concluding there can be no route to finding AI evidence inadmissible via hearsay.

Yet this viewpoint ignores the human behind the machine, which inherently complicates the question of whether AI can assert in a way simple calculators cannot. Scientists have written extensively on the issue of computer assertions in this new age of AI, cautioning that it is not the AI but "the humans who employ such systems who are responsible and sanctionable for the outputs and their effects."[218] Excluding AI from the protections of FRE 801 is a dangerous precedent, and the number of courts which have done just that demonstrates the need for an updated system of analyzing computer evidence. As another law professor argued, the "lack of understanding as to how" modern AI makes decisions will result in admitting unreliable statements because "the testimonial risks that are inherent in statements made by modern AI Entities are more akin to those found in human

[213] *Id.* at 88–89.

[214] FED. R. EVID. 801(a)–(c)(2).

[215] *Id.* at 801(a).

[216] *See* Grimm, Grossman & Cormack, *supra* note 176, at 85–86.

[217] United States v. Washington, 498 F.3d 225, 231 (4th Cir. 2007).

[218] Patrick Butlin & Emanuel Viebahn, *AI Assertion*, 2023 ERGO: OPEN ACCESS J. PHILOSOPHY, at 1, 24.

assertions that render them hearsay."[219] Providing blanket immunity from hearsay analysis to all types of machine output would allow individuals to have a computer "assert" on their behalf while being safe in the knowledge it will be admitted as objective evidence.

While amending the standards of admissibility would be a complicated and lengthy process, relying on the same rules that allow potentially faulty AI evidence to enter trials unquestioned and unimpeded is just as problematic. Ultimately, a problem cannot be solved using the same methods which created it, and AI admissibility requires a new way of thinking and a new, uniform answer.

V. A NEW SOLUTION

A. A Federal Agency and Court-Appointed Advisors

1. Filling in the Gaps of Previously Proposed Solutions

A review of the current literature on AI evidence quickly reveals there is little to no consensus regarding how to deal with this burgeoning problem, nor which organization should be responsible for doing so. If computer-driven evidence is to become the norm, there must be a recognizable standard across the court system for how to address it in order to preserve fair and consistent justice. A federal agency would do just that. Unlike potential rule revisions that fail to address every aspect of the problem or only serve to create more confusion, the solution of expert advisors organized by a centralized power would not suffer from those same drawbacks. And rather than leaving the situation alone and hoping these issues will resolve themselves over time, this provides a real mechanism for beneficial change.

Two pervasive issues plaguing each previous solution—ones that were even acknowledged by multiple authors—are the fact judges generally lack the training and experience necessary to understand computer programming, and black box programs preclude the possibility of understanding entirely. Usually, this is resolved by offering expert testimony at trial, but as discussed above, it is not as straightforward a process when it comes to AI, particularly AI operating via hidden source codes. Having a mechanism already in place to answer judges' questions about

[219] Jess Hutto-Schultz, *Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule*, 27 GEO. MASON L. REV. 683, 685 (2020).

machine evidence, one which is subject to strict rules of confidentiality, solves both issues.

Rather than forcing defendants to foot the bill of expensive experts, advisors sent by the agency would already be trained and available for use, reducing costs and increasing court efficiency. And instead of each party having to comply with complicated non-disclosure agreements, assuming companies agreed to release programming materials to them at all, a mechanism would already be in place to facilitate cooperation between those companies and the court system. By subjecting the agency and its advisors to strict confidentiality, companies' concerns of proprietary interests would be assuaged, more so than if an individual expert with potentially thievish motivations was given access to source codes.[220] A centralized agency would also address the fact that expert witnesses hired by a particular side carry an inherent risk of the witness being biased for that side.[221]

This solution aims not to revise the current admissibility rules that have proven to be more than adequate in other areas of evidence, but to provide an additional safeguard on top of the existing system. As explained by one researcher: "To expect competing for-profit companies to refrain from overclaiming and to fully disclose all uncertainties surrounding their findings is apparently expecting too much. To expect courts to regulate these matters as part of their review of admissibility apparently is also expecting too much."[222] The best solution is to filter machine evidence through an extra hurdle to ensure juries are not relying on faulty or misleading evidence represented as infallible truth. Defense counsels will not be forced to take a crash course on computer science to even hope to understand the technical documents behind a program's function. And judges' inability to

[220] Though individuals can seek information from federal agencies under the Freedom of Information Act, it specifically carves out an exception for "trade secrets and commercial or financial information" that is designed "to protect the interests of both the government and submitters of information." *FOIA Guide, 2004 Edition: Exemption 4*, U.S. DEP'T OF JUST., https://www.justice.gov/archives/oip/foia-guide-2004-edition-exemption-4 [https://perma.cc/9EE5-8D23] (Dec. 3, 2021).

[221] *See* Itiel E. Dror, Bridget M. McCormack & Jules Epstein, *Cognitive Bias and Its Impact on Expert Witnesses and the Court*, 54 JUDGES' J. 8, 9 (2015) ("[E]xperts are most often recruited by one side of the adversarial system, and work within the team and objectives of that side . . . [which] can subconsciously influence [the expert's] perceptions and judgments.").

[222] Thompson, *supra* note 149, at 1522.

research independently will no longer be an impediment to their valid exercise of discretion.

2. Analogous Positions Already in Existence

A scientific advisor is not a wholly novel or radical concept, nor is the creation of a centralized agency to ensure consistent administration of justice through those advisors. Judicial counsel positions are prevalent in courts to assist with conducting research for judges, family law courts often appoint psychologists to evaluate a child's mental state,[223] and law firms working in intellectual property generally hire scientific advisors for patent cases.[224] Despite the recent political trend of downsizing federal agencies,[225] the United States still has an abundance of them,[226] each organized around one central purpose. With the growing prevalence of AI in all facets of life, it seems nearly inevitable that a new agency will need to be created to address the accompanying concerns.

Both the United Kingdom[227] and India[228] have scientific advisors built into their intellectual property and patent law

[223] *The Judge Appointed a Psychologist in My Divorce Case. Now What?*, DADVOCACY (May 31, 2022), https://dadvocacy.com/blog/2022/05/the-judge-appointed-a-psychologist-in-my-divorce-case-now-what/ [https://perma.cc/JF74-KR33].

[224] *See* Lisa Larrimore Ouellette, *Transitioning from Science to Patent Law*, WRITTEN DESCRIPTION (Mar. 15, 2015), https://writtendescription.blogspot.com/2015/03/transitioning-from-science-to-patent-law.html [https://perma.cc/PC2P-8YK7].

[225] *See* Elena Shao & Ashley Wu, *The Federal Work Force Cuts So Far, Agency by Agency*, N.Y. TIMES, https://www.nytimes.com/interactive/2025/03/28/us/politics/trump-doge-federal-job-cuts.html [https://perma.cc/P23T-TQ4S] (May 12, 2025).

[226] *See The Federal Bureaucracy*, ADELPHI UNIV., https://libguides.adelphi.edu/c.php?g=745658&p=9242744 [https://perma.cc/M468-MYQT] (Aug. 23, 2024) (explaining that there are over 2,000 federal agencies in the United States which, together with the Cabinet departments, employ more than 2.7 million people).

[227] *See* Angus Milne & James Simpson, *Patents Court Provides Guidance on Technical Experts vs Scientific Advisers*, HLK (Apr. 4, 2024), https://www.hlk-ip.com/news-and-insights/patents-court-provides-guidance-on-technical-experts-vs-scientific-advisers/ [https://perma.cc/E345-GYBU]. In the United Kingdom, most scientific advisors are appointed to appeals courts, and they are meant to "educate the Court in the relevant technology" that is being offered as evidence, not assist in the determination of any substantive issue. *Id.*

> [T]here are cases at the cutting edge of science, where even the most experienced judges have considered that a non-controversial "teach-in" would be desirable and so the Court . . . has appointed a scientific adviser to assist it with getting to grips with the relevant concepts. Such teach-ins have been more commonplace in the Court of Appeal where it is almost inevitable that at least one [judge] . . . will not be steeped in the relevant science.

Brian Cordery, *The Role of Scientific Advisers in the English Patents Court*, KLUWER PAT. BLOG (Mar. 14, 2024), https://patentblog.kluweriplaw.com/2024/03/14/the-role-of-scientific-advisers-in-the-english-patents-court/ [https://perma.cc/7KNE-WSC3].

system as real-world demonstrations of how such a mechanism can effectively function. The simplest analogy in the United States is court-appointed psychologists in family court. Known as "evaluators," these licensed and trained professionals are ordered by a judge to assess the conditions of a home, determine which custody scenario would be in the best interest of the child, and provide a confidential report.[229] The final custody determination is made by the judge, but the evaluator provides vital information based on their years of experience in the field of mental health.[230] Similarly, the scientific advisor sent by the agency would offer expertise in how to evaluate the validity of various types of machine evidence, but the final decision on admissibility is still reliant on the judge's discretion. And while a psychological evaluator can be cost prohibitive since the parties must bear the expense themselves,[231] the process of providing scientific advisors would already be a component of the legal system available for use.

The cost-prohibitive nature of outside advisors has already negatively impacted defense teams in the context of AI. In the rare case a judge orders disclosure of TrueAllele's source code, Cybergenetics does not make it easy for defendants. As explained by one defense attorney, his client would have had to

[228] *See* Essenese Obhan & Sayali Gulve, *India: Appointing Scientific Advisors in Patent Disputes*, MONDAQ (Aug. 7, 2020), https://www.mondaq.com/india/patent/973846/appointing-scientific-advisors-in-patent-disputes [https://perma.cc/3VKH-CLAM].

> In India, Section 115 of the Patents Act . . . provides that in any suit for infringement or in any proceeding before a court, the court may at any time . . . appoint an independent scientific adviser, to assist the court or to inquire and report upon any such question of fact or of opinion . . . .
>
> . . . .
>
> A scientific advisor plays a crucial role in educating and presenting intricate technological issues [by] help[ing to] translate complex technology and communicate the legal implications of conclusions into terms the judges and the patent attorneys can understand.

Scientific Advisers in Patent Litigation – The Indian Perspective, IPR STUDIOS, http://iprstudio.com/scientific-advisers-in-patent-litigation-the-indian-perspective/ [https://perma.cc/7AQB-PFC3] (last visited Apr. 18, 2025).

[229] *California Courts Self-Help Guide: Child Custody Evaluations*, JUD. BRANCH CAL., https://selfhelp.courts.ca.gov/child-custody/evaluations [https://perma.cc/VQG4-LH4U] (last visited Apr. 9, 2025).

[230] *See id.*

[231] *See id.*; *see also* Joel S. Seidel et al., *What Is a 730 Custody Evaluation?*, JOEL S. SEIDAL & ASSOCS. (Mar. 15, 2017), https://www.seidellaw.com/blog/2017/march/what-is-a-730-custody-evaluation-and-what-should [https://perma.cc/6VQU-V66J] (explaining that a court-ordered evaluation "can cost anywhere from $1,000 to $100,000," depending on the issues to be assessed).

pay $15,000 to access and review the code.[232] On top of that staggering fee, "the defense expert would also have to obtain $1 million in liability insurance, agree to take only handwritten notes and travel to the company's Pittsburgh headquarters for the review."[233] In all, "it would cost at least $50,000 to comply with the nondisclosure agreement, which also might bar [the] expert witness from testifying."[234] By providing a nationwide mechanism, there will be a fairer system in which all parties can have equal access to AI expertise without going bankrupt.

B. How Would This Work?

While a proposed solution may seem viable or advantageous in the abstract, it must also survive the practicalities of operating within the real world. The central agency would likely need to have a broader purpose covering all areas in which AI is a growing issue, including intellectual property cases and law enforcement's use of AI, but one department would be focused on legal cases involving machine evidence. Parties could submit the evidence at issue to the department, and teams of vetted scientists would analyze the data, prepare a report, and send an advisor to counsel the judge as to the reliability of the evidence being presented. While an alternative method would simply be the use of court-appointed scientific advisors not connected to a centralized agency, concerns of inconsistent analysis and the inability to review millions of lines of code for potential errors without assistance would be assuaged by an agency's more collaborative, yet still confidential, environment.

Throughout the federal court system, there are 94 district courts,[235] and in the fiscal year 2023, there were 68,950 criminal defendant filings.[236] This means, on average, a district court handles over 730 criminal cases per year, discounting differences in jurisdiction that may result in a higher or lower caseload for particular courts. But not all of those cases will necessarily involve machine evidence. TrueAllele claims it has been used as

[232] Jouvenal, *supra* note 3.

[233] *Id.*

[234] *Id.*

[235] *Introduction to the Federal Court System*, U.S. DEP'T OF JUST., https://www.justice.gov/usao/justice-101/federal-courts [https://perma.cc/7W78-6LP6] (last visited Apr. 7, 2025).

[236] *Federal Judicial Caseload Statistics 2023*, U.S. CTS., https://www.uscourts.gov/statistics-reports/federal-judicial-caseload-statistics-2023 [https://perma.cc/APV3-BHQM] (last visited Apr. 8, 2025).

an analysis tool in over 1,000 cases to date,[237] and it was first used at trial in 2009.[238] Therefore, over the last fifteen years, TrueAlelle has been used as evidence at an average rate of 67 cases per year, which is only about 0.1% of those 68,950 yearly filings and less than a single case per district court. Assuming only one case per court, TrueAllele evidence would be present in just 0.1% of the 730 yearly cases. There are over 5,500 Assistant U.S. Attorneys (AUSAs) throughout the nation,[239] so with 68,950 filings in 2023, each AUSA handles on average 12 to 13 criminal cases per year, which constitutes about 2% of the 730 cases assigned to each court. If one AUSA is expected to handle 2% of the yearly caseload for a district court, a team of advisors would be more than capable of handling 0.1% of the same caseload, even given the more arduous and labor-intensive process of analyzing AI for potential falsehoods.

Assuming higher caseloads in certain jurisdictions, more cases involving AI after additional programs like COMPAS are taken into account, and higher rates of machine evidence moving into the future, the system should still be more than adequate. Using 2% as a standard caseload, there is a margin large enough to accommodate a 1,900% increase in the agency's caseload. Because the agency is not involved in substantive fact-finding but is designed to review program source codes as they pertain to particular evidence, the scientific advisors could easily move between multiple cases in a way that attorneys cannot. Furthermore, once the agency gains experience with a certain type of machine evidence and how its reliability should be weighed by judges, efficiency will naturally increase.

As a final safeguard to preserve the integrity of the justice system, the scientists hired by the agency would not only need a requisite degree and background in the scientific field at issue, but they must remain an impartial party not hired by either side nor by the company which owns the program being used. This would require a standardized vetting and certification process as well as the imposition of ethical requirements similar to those already imposed on officers of the court.

[237] *Demonstrating Our Expertise: Proven Technology*, CYBERGENETICS, https://www.cybgen.com [https://perma.cc/L5BS-HBAS] (last visited Apr. 8, 2025).

[238] Jouvenal, *supra* note 3.

[239] *NAAUSA Mission*, NAAUSA, https://www.naausa.org/about [https://perma.cc/GHB5-MK2W] (last visited Apr. 8, 2025).

VI. CONCLUSION

Though science fiction writers oft depict futuristic computer intelligences capable of answering questions humanity itself is incapable of solving, the current status of AI is somewhat less illustrious and more so fraught with errors. When those errors manifest in the court room, it places the entire system of justice at risk. As Richard Feynman—the famous American theoretical physicist and Nobel Laureate—once wrote, "What I cannot create, I do not understand."[240] AI can only mimic human creation; it cannot understand, and it cannot be trusted to make objective analyses uncontaminated by bias or human programmer error one hundred percent of the time.

The problem is not necessarily that machine evidence might not always be entirely accurate—courts face that same problem with human experts. The problem lies in the fact that jurors, and potentially even judges, more readily trust computer evidence without questioning it. Hidden algorithms, doubtful validation studies, and proven mistakes are chipping away at the foundation of the legal system. Without the institution of an extra measure of protection, cracks will only continue to form as AI gains further prominence as an evidentiary tool, eventually swallowing the court system whole and leaving defendants' fates in the hands of unseen algorithms. Then, computer programmers will really be "rolling their eyes."[241]

[240] *Richard Feynman's Blackboard at Time of His Death*, CALTECH ARCHIVES, https://digital.archives.caltech.edu/collections/Images/1.10-29/ [https://perma.cc/5TH6-SZVB] (last visited Apr. 9, 2025).

[241] *See* Baker, *supra* note 197 ("Behind every successful AI is a programmer rolling their eyes.").