SCIENTIFIC MODELS AND COMPUTER TIERS: A UNIFIED THEORY OF ADMISSIBILITY OF MACHINE-BASED EXPERT EVIDENCE

SCOTT DEVITO† & ANDREW W. JURS‡

Over the last few decades, researchers have challenged many routinely admissible types of evidence of a suspect's guilt or innocence (including analysis of bite marks, blood spatter, hair, shoe prints, and toolmarks) and found them lacking in foundational validity. Such methods were initially deemed revolutionary—providing new and more accurate methods of establishing guilt and innocence. Yet, their introduction likely caused many innocent persons to be found guilty and many guilty persons to be found innocent. The introduction of evidence from complex machine-based systems (including artificial intelligence (AI) computer programs) poses a similar challenge. It is one that is poised to provide more accurate ways of establishing guilt and innocence. And it is one that leads us down the same primrose path as bite marks, blood spatter, and hair analysis—to many innocents being found guilty and many of the guilty being found innocent. To help avoid the possibility of AI-based evidence misleading the courts, we propose a three-tiered framework for assessing the admissibility of expert testimony relying on machine-based evidence. Tier One evidence provides no new barriers to admissibility because the machine-based systems underlying it are transparent and explainable. Tier Two evidence arises from proprietary machine-based systems, making those systems only potentially transparent and explainable. As such, courts should admit evidence from Tier Two systems only with access to the underlying system (in the case of AI/computer programs, the code and data) so that it (and opposing counsel) can assess its reliability. Finally, courts should not admit Tier Three evidence, which is based on inherently opaque and unexplainable systems (like many neural networks), because the very structure of the systems creating the evidence makes it impossible for the court to assess its reliability. We contend that due process requires court-ordered disclosure of system architecture and examination by trained experts provided to the accused to ensure proper reliability

---

† Associate Professor of Law, Quinnipiac University School of Law; Ph.D. University of Rochester, J.D. University of Connecticut. The author thanks the Jacksonville University College of Law for its support and the Quinnipiac University School of Law for its continuing support.

‡ Robert Eglet Evidence Law Endowed Chair and Professor of Law, University of the Pacific, McGeorge School of Law. J.D., University of California, Berkeley, School of Law. Thanks to the McGeorge School of Law for its continued support. Thanks also to Katie, Clara, and Milo.

screening of AI-based evidence prior to admission. Significantly, we find constitutional protections under the Confrontation Clause to be unavailing because, on a practical level, although confrontation rights extend significantly beyond what courts require in the assessment of reliability, confrontation rights must be enforced through a reliability analysis.

Keywords: artificial intelligence, confrontation clause, *Daubert*, due process, evidence, expert witness, expert systems, neural networks

I. INTRODUCTION

An expert takes the stand to offer powerful evidence of the guilt of the accused in a criminal case. The defense responds by asking, "can you explain to me the steps in your process to reach your opinion?" The expert is unable to provide an answer. "Well, can you at least identify the factors most important to your decision?" The expert is unable to provide an answer. "At least, can you tell me why you think the opinion is correct?" The expert responds, "I was trained to do this, and in all my previous projects, I have never been found to be wrong!"

Such an expert should probably not clear their calendar for future court appearances, because any judge applying reliability gatekeeping standards would exclude the expert from testifying.[1] Reliability screening—and to some extent confrontation—requires an expert to be able to identify not only the conclusion they reach but also defend the methodology used to reach the conclusion.[2] The *Daubert* decision commands that when judges are doing gatekeeping, "[t]he focus . . . must be solely on principles and methodology, not on the conclusions that they generate";[3] moreover, the later *Daubert* trilogy cases, *Joiner* and *Kumho Tire*, both denounce *ipse dixit* expert opinions resting only on the say-so of the expert themself.[4]

Now imagine a second expert with the same opinion in the same case[5] testifies, but when asked about methodology, they simply point to reliance

1. Reliability gatekeeping is primarily an application of the *Daubert* standard from *Daubert v. Merrell-Dow Pharms.*, 509 U.S. 579 (1993), although some states continue to apply the older *Frye* standard of "general acceptance." Frye v. United States, 293 F. 1013 (D.C. Cir. 1923); People v. Leahy, 882 P.2d 321 (Cal. 1994); Donaldson v. Cent. Ill. Pub. Serv. Co., 767 N.E.2d 314 (Ill. 2002); Goeb v. Tharaldson, 615 N.W.2d 800 (Minn. 2000); People v. Wesley, 633 N.E.2d 451 (N.Y. 1994); Grady v. Frito-Lay, 839 A.2d 1038 (Pa. 2003); State v. Copeland, 922 P.2d 1304 (Wash. 1996).

2. *See* United States v. Frazier, 387 F.3d 1244, 1265–66 (11th Cir. 2004) (finding that trial court did not abuse its discretion by excluding expert witness testimony for which the witness failed to explain the basis for their opinion); Rodrigues v. Baxter Healthcare Corp., 567 F. App'x. 359, 361–62 (6th Cir. 2014) (finding that trial court did not abuse discretion by excluding testimony of expert witness who could not explain process underlying testimony); Johnson v. J.P. Parking, Inc., 717 F. Supp. 3d 798, 815–17 (S.D. Iowa 2024) (excluding testimony of expert witness who did not explain the factors used, quantify the factors, or describe how they led to his conclusion).

3. *Daubert*, 509 U.S. at 595.

4. Kumho Tire v. Carmichael, 526 U.S. 137, 157 (1999) (rejecting the expert in the case because the expert "himself claimed that his method was accurate, but, as we pointed out in *Joiner*, 'nothing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert'"); Gen. Elec. v. Joiner, 522 U.S. 136, 146 (1997) (finding that "nothing in either Daubert or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert").

5. *See* hypothetical case discussion *supra* p. 244.

on a computer system trained with AI. One would expect the evidence to meet the same fate as the first expert, as they would assume that a court would similarly exclude it under any form of reliability screening. But one would be wrong.[6] Courts admit expert testimony based on the output from AI systems (mainly in the area of DNA mixture analysis)[7] in criminal cases without any detailed assessment or even disclosure of the reliability of the method used to create the final conclusion.[8] Instead, courts have leaned on the programmer's internal validation studies in deciding to admit the opinions.[9]

This is the exact approach that has, in the past, led to disastrous consequences when courts admitted invalid, shoddy, or untested forensic evidence in criminal cases based on long-standing norms, judicial notice, other courts' acceptance of the method, or lack of serious inquiry into the method.[10] For example, the President's Council of Advisors reviewed

---

6. *See, e.g.*, United States v. Gissantaner, 990 F.3d 457, 463–66 (6th Cir. 2021) (finding that STRMix need only be capable of being tested, that peer review need not be from independent reviewers, and that it has been admitted by other courts); People v. Wakefield, 195 N.E.3d 19, 21, 30 (N.Y. 2022) (finding that the TrueAllele evidence was reliable enough to admit under New York's *Frye*-based general acceptance standard and that experts need not disclose the source code, due to the validation studies).

7. Regarding cases that admit expert testimony based on artificial intelligence systems with DNA mixture analysis, see *United States v. Anderson*, 673 F. Supp. 3d 671 (M.D. Pa. 2023) and *Wakefield*, 195 N.E.3d 19.

8. Cases that allow the evidence without disclosing the underlying code for the analysis of the DNA mixture far outnumber those where courts require disclosure. Courts have generally admitted probabilistic genotyping results without requiring disclosure of the underlying source code. *See, e.g.*, *Wakefield*, 195 N.E.3d at 26–27 (affirming trial court's conclusion that access to TrueAllele's source code was not required to establish reliability under *Frye*); People v. Superior Court (Dominguez), 28 Cal. App. 5th 223, 233, 243 (Cal. Ct. App. 2018) (rejecting order compelling STRmix source code and reversing trial court); People v. Davis, 75 Cal. App. 5th 694, 721–22 (Cal. Ct. App. 2022) (affirming admissibility of STRmix results despite defendant's inability to review black-box code); Commonwealth v. Foley, 38 A.3d 882, 889–90 (Pa. Super. Ct. 2012) (upholding TrueAllele evidence without full access to internal algorithms); People v. H.K., 69 Misc.3d 774, 784–85 (N.Y. Crim. Ct. 2020) (rejecting argument that confrontation and due process required disclosure of STRmix code). By contrast, the authors have found only one case in which the court compelled code disclosure. *See* State v. Pickett, 246 A.3d 279, 306–07 (N.J. App. Div. 2021) (ordering disclosure of TrueAllele source code under strict confidentiality conditions).

9. As to internal validation testing being the basis for admission, see *Gissantaner,* 990 F.3d at 467; *Anderson*, 673 F. Supp. 3d at 682–83; *Wakefield*, 195 N.E.3d at 28.

10. *See* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH.., EXEC. OFF. OF THE PRESIDENT, FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS 6–13 (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf [https://perma.cc/E59E-628W] (explaining and

scientific studies in the field of hair comparison evidence (once commonly admitted and, for decades, considered a major forensic field)[11] and found that those studies do not support the field's foundational validity.[12] Similarly, when examining their "scientific" underpinnings closely, researchers have found that bite marks, toolmarks, shoe prints, blood spatter, and a variety of other fields lack scientific merit.[13]

The growing power and prevalence of AI has placed the criminal justice system on the verge of another revolutionary period of change.[14] But it would be foolhardy for courts to throw caution to the wind and find AI-based evidence valid solely on the word of the developers. Instead, we propose that such systems should be available for a detailed examination of their system's architecture to ensure reliability of the system before admission in court. Such an approach provides a simple, clear standard and eliminates any confusion that could arise in the courts regarding when to admit such evidence.[15]

To support this proposal, we first examine the types of machine-based evidence that are admissible today, the current legal rules that apply to their admission in court, and how courts have addressed these fields.[16] As part of this process, we examine the implications of evidence based on

---

comprehensively reviewing why several forensic testing methods are flawed, including methods for microscopic hair comparison, bullet-lead evidence, footwear analysis, and bitemarks); NAT'L RSCH. COUNCIL, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 99–109 (2009), https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf [https://perma.cc/B6YP-8XKP] (reviewing the admission of faulty forensic analysis, from multiple disciplines, that occurs without detailed review of the method's underlying scientific validity).

11. *See* Ed Pilkington, *Thirty Years in Jail for a Single Hair: the FBI's 'Mass Disaster' of False Conviction*, THE GUARDIAN (Apr. 21, 2015), https://www.theguardian.com/us-news/2015/apr/21/fbi-jail-hair-mass-disaster-false-conviction [https://perma.cc/7WEW-C35S] (reporting that forensic hair examination began in 1855 and has since been frequently admitted as evidence of an accused's guilt).

12. *See* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 10, at 13.

13. *Id.* at 7–14.

14. *See* DEP'T OF JUST., ARTIFICIAL INTELLIGENCE AND CRIMINAL JUSTICE: FINAL REPORT (2024), https://www.justice.gov/olp/media/1381796/dl [https://perma.cc/9FFQ-3BFR] (reporting on the need for expanding AI use in the criminal justice system and the potential related implications).

15. *Compare* United States v. Gissantaner, 990 F.3d 457, 463–64 (6th Cir. 2021) (concluding that AI-based evidence was admissible without requiring examination of the underlying code), *and* People v. Wakefield, 195 N.E.3d 19, 21, 29 (N.Y. 2022) (same), *with* State v. Pickett, 246 A.3d 279, 317 (N.J. Super. Ct. App. Div. 2021) (finding that adversarial review of black box system's source code was appropriate), *and* United States v. Gissantaner, 417 F. Supp. 3d 857 (W.D. Mich. 2019), *rev'd*, 990 F.3d 457, 468 (6th Cir. 2021) (reversing, in effect, trial court's determination that access to the source code was necessary).

16. *See* discussion *infra* Part IV.

proprietary expert systems and closed-box neural networks.[17] This leads us to place machine-based evidence (including AI-based evidence) into three tiers based on the system's transparency and explainability.[18]

We then examine how judges should handle gatekeeping analyses with each tier having a different admissibility standard. Tier One systems are inherently transparent and explainable, and therefore are admissible,[19] while Tier Three systems are, by their design, inherently opaque and non-explainable, and therefore, properly excluded.[20] However, Tier Two systems are a more complex area because the systems are proprietary; as such, while they may be ***potentially*** explainable and transparent, they are not ***actually*** explainable and transparent because experts do not disclose their source code.[21] We therefore conclude that the only way for a court to engage in proper reliability screening is to require the disclosure of the underlying system code of Tier Two systems and then use that code to assess its reliability.[22]

Lastly, we examine constitutional standards that apply to AI-based expert evidence.[23] We find, perhaps counterintuitively, that current interpretations of the Confrontation Clause right do not provide meaningful protection independent of reliability screenings.[24] Instead, we believe that it is an accused's due process right to an expert—specifically one to assist in the examination of the source code disclosures for mixed

---

17. *See* discussion *infra* Sections IV.B, IV.C.
18. *See* discussion *infra* Part IV.
19. *See* discussion *infra* Section IV.A.
20. *See* discussion *infra* Section IV.C.
21. *See* discussion *infra* Section IV.B.
22. *See* discussion *infra* Section IV.B.
23. *See* discussion *infra* Section V.B.
24. We say counterintuitively because recent commentators in the field have examined the Confrontation Clause in detail as a procedural limitation on machine evidence. *See, e.g.*, Andrea Roth, *What Machines Can Teach Us About Confrontation*, 60 DUQ. L. REV. 210, 211–13 (2022) (reviewing current confrontation doctrine and arguing for a broader right to meaningfully scrutinize the government's proof, beyond witness questioning); Edward Cheng & Alex Nunn, *Beyond the Witness: Bringing A Process Perspective to Modern Evidence Law*, 97 TEX. L. REV. 1077, 1081–82 (2019) (reviewing current confrontation doctrine and arguing for a process based approach allowing reliability screening beyond witness examination); Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 2044–48 (2017) (arguing to include machine testimony within confrontation clause protections); Erin Murphy, *The Mismatch Between Twenty-First-Century Forensic Evidence and Our Antiquated Criminal Justice System*, 87 S. CAL. L. REV. 633 (2014) (reviewing current confrontation doctrine and arguing for a broader right to systematic transparency). Many of these articles recognize the limitations of the current confrontation doctrine with computer evidence and thus argue for a reinterpretation of the current doctrine to take this evidence into account.

models—that provides the more significant protection for the accused against AI-based expert evidence.[25]

The last few decades have been a time for the criminal justice system to reflect on past errors and reconsider long-standing beliefs.[26] Evidence once deemed routinely admissible has been challenged and found lacking in foundational validity.[27] We contend that due process requires disclosure of the systems creating AI-based evidence (and all machine-based evidence), as well as the provision of an expert (where necessary) who can assess the reliability of that system to ensure proper screening of the evidence.

II. THE LAW OF RELIABILITY AND CONFRONTATION IN MACHINE-BASED SCIENTIFIC FORENSIC ANALYSIS

Complex scientific testimony based on machine analysis is not a new issue in the judicial system, as experts have relied on tools to assess evidence since a microscope was first aimed at a fingerprint, if not earlier.[28] Yet the significant advances in AI suggest that forensic analysis performed largely by, or even exclusively by, AI systems is in the not-too-distant future.[29] So how would a court handle the decision to admit or exclude an AI-based forensic analysis in those cases? A review of case law examining machine-assisted and computer-assisted evidence helps demonstrate the issues judges must assess in these circumstances, and how

---

25. The due process right to an expert in *Ake v. Oklahoma* provides a "fair opportunity to present [a] defense," and so the state will provide an expert, at state expense, when the court deems that expert to be one of the "raw materials integral to the building of an effective defense." Ake v. Oklahoma, 470 U.S. 68, 76–77 (1985). Regarding the wide variety of experts that courts have found to be necessary as a matter of due process, see *infra* note 301 and accompanying text.

26. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH, *supra* note 10, at 13.

27. *See, e.g.*, *id.* at 7–14 (discussing validity concerns with widely accepted forensic methods).

28. The National Research Council report, *Strengthening Forensic Evidence in the United States: A Path Forward*, identifies the first published case in the United States regarding fingerprints as *People v. Jennings*, a 1911 case from Illinois. NAT'L RSCH. COUNCIL, *supra* note 10, at 72–73 (reviewing the history of fingerprint evidence in the late 19th and early 20th Centuries).

29. 1 PAUL GIANNELLI ET AL., SCIENTIFIC EVIDENCE PREFACE (6th ed. 2020) ("Even as of this writing, advances in machine learning and deep neural networks suggest that, by the [next] Edition of this treatise, much of scientific expert "testimony" may be delegated to artificial intelligence."). For a review of the (then) current and possible future use of artificial intelligence in forensics, see Andrew W. Jurs & Scott DeVito, *Machines Like Me: A Proposal on the Admissibility of Artificially Intelligent Expert Testimony*, 51 PEPP. L. REV. 591, 604–09 (2024).

current evidentiary frameworks and Confrontation Clause doctrine would apply to AI-based evidence.

A. Reliability in Scientific Forensic Analysis

Courts can admit evidence based on "scientific, technical, or other specialized knowledge" through Federal Rule of Evidence (FRE) 702, but the judge has the obligation to review the evidence for relevance and reliability prior to admission.[30] Reliability analysis can be a complex, difficult task, and depends on the judge assessing the underlying methodology of the proffered testimony to make sure it is good science.[31] The *Daubert* decision states unequivocally that the "focus, of course, must be solely on principles and methodology, not on the conclusions that they generate."[32]

Judges can use any method they wish to make these reliability determinations, although many will hold an evidentiary hearing, sometimes with live testimony, to assess if the methodology is "good science."[33] On a substantive level, the judge may wish to consider a series of factors listed by the *Daubert* court—testing, peer review and publication, standards, error rate, and general acceptance[34]—to ensure the method is sound, although they are not limited solely to these considerations particularly when addressing non-lab coat technical expertise.[35] The court has cautioned judges to ensure that there is not "too great an analytical gap between the data and the opinion proffered."[36]

B. Confrontation in Scientific Forensic Analysis

For those cases that involve criminal charges, there is an additional consideration: the right to confrontation. The Sixth Amendment's Confrontation Clause provides that "[i]n all criminal prosecutions, the accused shall enjoy the right . . . to be confronted with the witnesses against him."[37] Generally this means that a criminal defendant has the right

30. Daubert v. Merrell Dow Pharms., 509 U.S. 579, 589 (1993) (emphasis omitted).
31. *Id.* at 590–93.
32. *Id.* at 595.
33. Andrew W. Jurs, *Gatekeeper with a Gavel: A Survey on Judicial Management of Challenges to Expert Reliability and Their Relationship to Summary Judgment*, 83 MISS. L.J. 325, 332–33, 347 (2014) (discussing prior research on methods used to resolve reliability challenges, then reviewing results of a new survey).
34. *Daubert*, 509 U.S. at 593–94.
35. Kumho Tire v. Carmichael, 526 U.S. 137 (1999).
36. Gen. Elec. Co. v. Joiner, 522 U.S. 136, 146 (1997).
37. U.S. CONST. amend. VI.

to ask questions of a witness offered by the prosecution, in cross examination, to allow the jury to assess the witness's demeanor, memory, or overall credibility.[38] Since 2004, the Court has interpreted the right of confrontation to apply solely to statements that are "testimonial" in nature, which the *Crawford v. Washington*[39] decision found applied to court testimony and formal police statements.[40] A decade ago, in *Melendez-Diaz v. Massachusetts*[41] and *Bullcoming v. New Mexico*,[42] the Court addressed confrontation with expert analysis, finding that the Confrontation Clause does apply to scientific analysis of evidence and, consequently, the analyst who performed a specific lab test must testify (and be subject to cross examination) to admit the results of that test in a criminal case.[43]

One question the Supreme Court struggled to decide was whether the basis of expert testimony is testimonial, therefore triggering a defendant's confrontation right. In *Williams v. Illinois*,[44] the court evaluated the testimony of a DNA technician who concluded that the sample from the defendant examined at the crime lab matched a profile from a private company, Cellmark, based on a swab taken from the crime victim.[45] The Court decided, in a plurality opinion, that when a scientist relies upon materials to form an expert opinion, the underlying information is not testimonial because courts do not consider it for the truth (only for basis of opinion) and, if so, the Confrontation Clause would be inapplicable.[46] For many years after the Court decided *Williams* in 2012, courts struggled to apply it to a variety of other situations.

Recently, however, the Supreme Court offered clarity on the issue addressed in *Williams*, in the 2024 decision of *Smith v. Arizona*.[47] In *Smith*,

---

38. *See, e.g.*, United States v. Owens, 484 U.S. 554, 559 (1988) (finding confrontation has been satisfied when the defendant "has the opportunity to bring out such matters as the witness' bias, his lack of care and attentiveness, his poor eyesight, and even . . . the very fact that he has a bad memory"); California v. Green, 399 U.S. 149, 159 (1970) (rejecting a Confrontation Clause appeal since the witness was subject to cross for the jury, who could observe his demeanor and assess credibility of the witness).
39. Crawford v. Washington, 541 U.S. 36 (2004).
40. *Id.* at 68. *Davis v. Washington*, two years later, would expand on the concept by defining "testimonial" using a "primary purpose" test, which remains in effect today. Davis v. Washington, 547 U.S. 813, 822 (2006).
41. Melendez-Diaz v. Massachusetts, 557 U.S. 305 (2009).
42. Bullcoming v. New Mexico, 564 U.S. 647 (2011).
43. *Id.* at 652 (requiring the analyst who performed the test to testify in order to satisfy the Confrontation rights of the accused); Melendez-Diaz, 557 U.S. at 324 (rejecting claim that affidavit showing results of chemical test satisfied Confrontation rights of the accused).
44. Williams v. Illinois, 567 U.S. 50 (2012), *abrogated by*, Smith v. Arizona, 602 U.S. 779 (2024)
45. *Id.* at 56.
46. *Id.* at 57–58.
47. Smith v. Arizona, 602 U.S. 779 (2024).

the prosecution alleged he was in possession of illegal narcotics and, to prove those charges, needed to present evidence of the chemical tests of the substances seized on the accused's property.[48] Initially, the prosecution endorsed the lab technician who had run the drug tests at the crime lab, but by the time of the trial, that technician was unavailable, so they endorsed a second lab analyst who concluded the seized substances were unlawful narcotics.[49] In the lower courts, judges relied on *Williams* to conclude that if the second analyst used the first analyst's notes as a proper basis of opinion under FRE 703, then the analyst did not offer them for the truth, and, thus, there is no confrontation protection.[50] Unlike in *Williams*, however, the *Smith* court offered a unanimous ruling with a clear rule moving forward, that "[i]f an expert for the prosecution conveys an out-of-court statement in support of his opinion, and the statement supports that opinion only if true, then the statement has been offered for the truth of what it asserts."[51] So if the testimony of the second expert relied on the chemical test of the first expert, and it could only form a proper basis for testimony if that test result is true, then the statement of the first expert has been offered for the truth.[52] As of this writing, of course, few cases have been able to apply the *Smith* holding, but it does offer clarity on a point previously unclear in the law and provides a clear implication for AI-based expert testimony as discussed in detail below.[53]

III. A NOTE ON TERMINOLOGY

When discussing scientific expert testimony, the courts generally reference the scientific "theory" supporting the expert's opinion.[54] In our discussion below, we use the term "model" in place of "theory" for three reasons. First, theories describe abstract entities in idealized conditions—not real systems and objects.[55] For example, "[a] simple pendulum is a mass swinging from a massless string attached to a frictionless pivot, subject to a uniform gravitational force, and in an environment with no resistance. This is clearly an ideal object. No real pendulum exactly

---

48. *Id.* at 789–90.
49. *Id.* at 790–91.
50. *Id.* at 791–92.
51. *Id.* at 795.
52. *Id.* at 796–99.
53. See discussion *infra* Sections IV.A.3, IV.A.4, and V.B.1.
54. *See, e.g.*, Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593 (1993); L.M. *ex rel.* Dussault v. Hamilton, 436 P.3d 803, 811 (Wash. 2019) (discussing the *Frye* standard).
55. *See* FREDERICK SUPPE, THE SEMANTIC CONCEPTION OF THEORIES AND SCIENTIFIC REALISM 83 (1989).

satisfies any of these conditions."[56] As a result, scientific experts do not directly rely upon a theory—they must use a model that instantiates the theory relative to the real objects and forces that are the subject of the opinion.

Second, because theories relate to abstract entities, not real ones, the same theory can generate different models with different objects and conditions.[57] For example, Euclid's theory of geometry spawns models including spatial relationships in computer graphics,[58] preferences in economics,[59] and biological relationships.[60] As a result, referencing just the theory is insufficient to understand what the scientist is actually using—the associated model.

Finally, this article's central concern is with the introduction of the results from machine-learning algorithms as machine-based evidence. Practitioners and theoreticians of machine-learning describe those algorithms as producing *models* of the data.[61] Thus, the use of "model," as opposed to "theory," is also consistent with the understanding and usage of persons working in machine learning.

IV. THREE TIERS OF MACHINE-BASED SCIENTIFIC FORENSIC ANALYSIS

Courts must base scientific expert testimony on reliable principles and methodologies supported by reference to scientific models.[62] As we show below,[63] these constraints require that machine-based evidence be both (1) transparent[64] and (2) explainable.[65] At the same time, while criminal

56. RONALD GIERE, SCIENCE WITHOUT LAWS 122 (1999).
57. *See, e.g.*, 37 OWEN BYER ET AL., METHODS FOR EUCLIDEAN GEOMETRY 10–11 (1st ed. 2010) (discussing concrete and abstract models for Euclidean geometry).
58. *See, e.g.*, David Fontijne and Leo Dorst, *Modeling 3D Euclidean Geometry*, 23 IEEE COMP. GRAPHICS & APP. 68–78 (2003).
59. *See, e.g.*, Anna Bogomolnaia & Jean-François Laslier, *Euclidean Preferences*, 43 J. MATH. ECON. 87–98 (2007).
60. *See, e.g.*, Mark Layer & John A. Rhodes, *Phylogenetic Trees and Euclidean Embeddings*, 74 J. MATH. BIOL. 99–111 (2017).
61. *See, e.g.*, ETHEM ALPAYDIN, INTRODUCTION TO MACHINE LEARNING 1 (4th ed. 2020); Roar Arne Fjellheim, *Model-Based Reasoning*, *in* THE HANDBOOK OF APPLIED EXPERT SYSTEMS 9-1–9-15 (Jay Liebowitz ed., 1998).
62. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 595 (1993).
63. *See* discussion *infra* Part IV.
64. A model is transparent if it is sufficiently open and accessible to review. *See, e.g.*, Adrian Weller, *Transparency: Motivations and Challenges*, *in* EXPLAINABLE AI: INTERPRETING, EXPLAINING AND VISUALIZING DEEP LEARNING 23, 23–27 (Wojciech Samek et al. eds., 2019).
65. A model is explainable if there is sufficient information about the model to enable the court to understand the connection between the model's input and its output. *See, e.g.*,

defendants have a right to confront testimonial witnesses, the practical application of the confrontation limit makes it unlikely that confrontation provides an independent barrier to admission as any machine-based evidence with Confrontation Clause concerns would also be unreliable under *Daubert*.[66] Nonetheless, as we argue, due process imposes limitations on the admissibility of certain kinds of complex machine-based evidence and requires the court to provide computer forensic experts to aid the defense to ensure that the system is both transparent and explainable.[67]

Scientific experts can rely on machine-based evidence from scientific instruments because those instruments are built in accord with a known (*transparent*) underlying scientific model that *explains* how the instrument works and why it produces the results that it does.[68] But, as discussed below, as the explainability of a model decreases, so does our ability to judge its reliability.[69] This feature of machine-based evidence creates three tiers of machine-based evidence based on the degree of transparency and explainability of the machine-system.

Tier One systems, including gas chromatography, mass spectrometry, and computer-assisted DNA electrophoresis, generally satisfy reliability and confrontation because the underlying scientific models are transparent and explainable.[70] Such systems provide no special impediments to admissibility.

Tier Two systems such as DNA mixture analysis rely on complex computer algorithms[71] including expert systems[72] (a form of AI[73]) as integral parts of the data-generation process. Because those algorithms simply function as a different kind of model, they are intrinsically no different from Tier One systems. But because computer algorithms may

---

Marco Tulio Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, *in* PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1136 (2016).

66. See discussion *infra* Sections IV.2, IV.4, and V.B.1.

67. *See* discussion *infra* Section V.B.2.

68. For a detailed discussion of the need for theory underlying scientific instruments see IAN HACKING, REPRESENTING AND INTERVENING 186–209 (1983) (discussing microscopes).

69. *See* discussion *infra* Part IV.

70. See discussion *infra* Sections IV.A.2, A.5, and B.2–3.

71. We define "algorithm" as a process or set of rules to be followed in calculations or problem solving. *See, e.g.*, Allen Clark Zoracki, *When is an Algorithm Invented? The Need for a New Paradigm for Evaluation*, 15 ALBANY L. J. SCI. & TECH. 579, 581 (2005).

72. *See* MARK W. PERLIN, AN EXPERT SYSTEM FOR SCORING DNA DATABASE PROFILES, https://www.promega.com/~/media/files/resources/conference%20proceedings/ishi%2011/oral%20presentations/perlin.pdf [https://perma.cc/WP97-3L5E] (last visited Dec. 22, 2025) (providing a press release relating to the use of expert systems in the TrueAllele system).

73. *See, e.g.*, ITISHA GUPTA & GARIMA NAGPAL, ARTIFICIAL INTELLIGENCE AND EXPERT SYSTEMS 24 (2020).

be extremely complex and proprietary, Tier Two systems are only *potentially* transparent and explainable.[74] Consequently, expert testimony that relies on Tier Two machine-based evidence should only be admitted if the systems are *actually* transparent and explainable.[75]

Finally, Tier Three systems are machine-based systems that are intrinsically closed-box systems in that they are neither transparent nor explainable.[76] At present, many neural network systems are closed-box systems in that their creation makes it impossible for users to understand the model used by the neural network and, as a consequence, makes it impossible to understand or explain why the system produced any particular result.[77] While neural networks and other forms of machine learning are only currently found in the law as drafting assistance tools, document analysis (e.g., e-discovery), text retrieval, case/legislation analysis, speech-to-text, chatbots, and office administration,[78] given the growing prevalence and power of large-language models (LLMs) like OpenAI's ChatGPT (a closed-box system[79]), we should expect efforts to introduce closed-box machine-learning evidence in the very near future. We argue below that, due to closed-box systems' lack of transparency and explainability, evidence based on the output from closed-box systems should not be admitted because it is impossible to assess its reliability, and admission would violate criminal defendants' due process rights.[80]

A. Tier One Systems: Simple and Knowable Models

Imagine a drug case where the police find multiple baggies of a powder substance on the driver after an arrest for a minor traffic violation,[81] or after a suspicious fire, the investigators want to decide if there is an accelerant present and, if so, which one.[82] Imagine a different case where a DNA sample is taken from a crime scene, and the crime lab wishes to assess whether it matches a known sample taken from a suspect.

---

74. *See* discussion *infra* Section IV.B.
75. *See* discussion *infra* Section IV.B.
76. *See* discussion *infra* Section IV.C.
77. *See, e.g.*, Mindy Duffourc & Sara Gerke, *Decoding U.S. Tort Liability in Healthcare's Black-Box AI Era: Lessons from the European Union*, 27 STAN. TECH. L. REV. 11–12 (2024) (discussing black-box neural networks).
78. *See, e.g.*, PETER HOMOKI, EUR. LAWS. FOUND., GUIDE ON THE USE OF ARTIFICIAL INTELLIGENCE-BASED TOOLS BY LAWYERS AND LAW FIRMS IN THE EU 20–37 (2022).
79. *See* Matthew Kosinski, *What is Black Box Artificial Intelligence (AI)?*, IBM (Oct. 29, 2024), https://www.ibm.com/think/topics/black-box-ai [https://perma.cc/6TB4-3Y58] (noting that ChatGPT is a black-box system).
80. See discussion *infra* Sections IV.C, V.A, and V.B.2.
81. State v. Klein, 679 S.W.3d 147 (Mo. Ct. App. 2023).
82. United States v. Aman, 748 F. Supp. 2d 531 (E.D. Va. 2010).

In each case, the prosecution will offer evidence to connect the evidence collected to the suspect to help prove the crime. To make that connection, the police may, in the first case, decide to rely on chemical analysis done by a gas chromatograph and mass spectrometer, and in the second, use computer-assisted electrophoresis to decide if there is a match.

In both areas, the type of machine-based evidence will meet any standards for admissibility because each step of the process is based upon known scientific models that are transparent and explainable.[83]

1. Gas Chromatography and Mass Spectrometry

To understand how courts assess the reliability of machine-assisted chemical analysis, it is necessary to understand the mechanism by which the machine makes a determination of the content of an unknown sample. This involves a two-step process. In Step One (vaporization), the sample is vaporized into components that can be contained by an inert gas within the machine.[84] In Step Two (detection), the components of vaporization move through a capillary to be read by a detector (often a mass spectrometer), which determines their molecular structure and the quantity of each molecule.[85]

There are multiple steps necessary for proper vaporization of the unknown sample at Step One. The gas chromatograph device consists of purified silica tubes, ten to one hundred meters long, each with an inside diameter of 0.1 to 0.5 mm.[86] These tubes are coiled within an oven structure, which allows for temperature control within the capillary tubes.[87] Each tube has a coating of an immobilized liquid, the stationary phase, which will interact with the samples when they are injected into the system.[88] When a sample of an unknown substance is injected into the system, the compound is separated based on the interactions with the stationary phase, and the movement of the separated components through the system depends on the magnitude of the chemical interactions between the individual ion and the stationary phase.[89] Eventually, all components

83. *See* discussion *infra* Sections IV.A.1–2; IV.A.4–5.
84. JOHN W. MOORE ET AL., CHEMISTRY: THE MOLECULAR SCIENCE 544 (4th ed. 2011). For a description of the GC/MS process, see generally DAVID FAIGMAN ET AL., 4 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY § 40:36; 2 PAUL GIANNELLI ET AL., SCIENTIFIC EVIDENCE §23:03[4] (6th ed. 2020).
85. MOORE ET AL., *supra* note 84, at 544. For a description of the GC/MS process, see generally FAIGMAN ET AL., *supra* note 84, § 40:36; PAUL GIANNELLI ET AL., *supra* note 84, §23:03[4].
86. MOORE ET AL., *supra* note 84, at 544.
87. *Id.*
88. *Id.*
89. *Id.*

will reach the detector, but the time it takes to reach the detector will vary depending on the component in question.[90] The more interaction between the stationary phase and an individual component of the compound, the longer the ion remains within the capillary column, and conversely, the less interaction of the compound with the stationary phase, the sooner the ion is expelled from the capillary column into the detector.[91] Eventually, all individual components will be expelled from the column to be read by the detector at the end of the system, through a process called "differential migration."[92]

Once the components of the unknown compound reach the detector (Step Two), the detector records the time of the ejection of the compound and the amount of the substance in that component.[93] A mass spectrometer is a common detection method used in chemical analysis, one that the National Research Council describes as the "near universal test for identifying unknown substances."[94] Within the mass spectrometer, the components entering from the gas chromatograph flood with electrons, resulting in collisions between the electrons and the compound components.[95] This creates positive ions, which, due to their positive charge, are attracted to a negatively charged grid.[96] At this point, the ions pass through a magnetic grid, which deflects the ions at different rates depending on the size of the ions in question; larger ions deflect less while smaller ones deflect more, and this allows the machine to sort the ions by mass.[97] The resulting readout will demonstrate the abundance of each ion and its weight, leading to a graph output called a "mass spectrum."[98] The spectrum is a graph that shows all components of the unknown sample, and that pattern will be unique to different chemical compounds.[99] Thus, a mass spectrum of gasoline will have one particular signature pattern, while methamphetamine will have another. Taking the resulting pattern

---

90. *Id.*
91. *Id.*
92. *See, e.g.*, H.W. Paton, *Fundamental Principles*, *in* PRINCIPLES AND PRACTICE OF GAS CHROMATOGRAPHY 8–10 (Robert L. Pecsok ed. 1959); Richey v. Mitchell, 395 F.3d 660, 666 (2005) (discussing differential migration).
93. MOORE ET AL., *supra* note 84, at 544.
94. NAT'L RSCH. COUNCIL, *supra* note 10, at 135 (2009).
95. MOORE ET AL., *supra* note 84, at 56.
96. *Id.*
97. *Id.*
98. *Id.*
99. PAUL GIANNELLI ET AL., *supra* note 84, §23:03, nn.219–20, §23.03[4][c][i] ("The pattern is unique to the compound.").

and comparing it to known tables of common chemical compounds should result in a positive identification of the sample.[100]

While this system is complex, it is also transparent and explainable because, as we have demonstrated above, each of its steps can be explained by using standard, well-known scientific models like differential migration.

2. The Gas Chromatography and Mass Spectrometry Technique Is Reliable

While the gas chromatography and mass spectrometry (GC/MS) technique has been repeatedly described as the "gold standard" for chemical analysis, it does have limitations that can affect the accuracy of the final analysis.[101] First, the analyst operating the machine must be fully trained, and the machine must be operated correctly to reach a reliable result.[102] Next, the analyst must compare the resulting mass spectrum to known references to determine the identity of the substance.[103] Professor Paul Giannelli, in the reference manual *Scientific Evidence*, notes that the identification "is like working a jigsaw puzzle," and can lead to error through interpretation.[104] Even if using computer software for assessment, the analysis in determining the unknown substance ultimately is a comparison not unlike fingerprint analysis or other similar forensic fields.[105] Finally, there is an inherent limitation in identifying substances that are made up of the same molecular components but differ in the spatial arrangement of those components; for example, gas chromatography and mass spectrometry cannot distinguish ephedrine from pseudoephedrine because they have the same chemical makeup but differ solely in the spatial arrangement of the hydroxyl group.[106]

Yet even with these weaknesses, judges regularly admit the GC/MS method to determine the content of unknown substances in court. Having reviewed the methodology of the analysis, it is not hard to see why

100. *Id.* at n.224 and accompanying text (describing the method of analysis to compare the spectrum to known reference guides); *see also* FAIGMAN ET AL., *supra* note 84, § 40:36 ("The MS fragmentation pattern is unique for any given drug.").

101. FAIGMAN ET AL., *supra* note 84.

102. *Id.* § 40:36 ("Interpretation of GC, and particularly GC/MS data, requires technical expertise that can only be gained through years of training and instrument operation.").

103. PAUL GIANNELLI ET AL., *supra* note 84, §23:03[4][d].

104. *Id.*

105. *Id.* at n.235.2 (citing *How Good Are the Data? Novel Metric Assess Probability That an Unknown Drug Sample Matches a Known Sample*, NAT'L INST. JUST. (Mar. 17, 2022), https://nij.ojp.gov/topics/articles/how-good-are-data-novel-metric-assesses-probability-unknown-drug-sample-matches [https://perma.cc/E7YH-ZMNK]).

106. PAUL GIANNELLI ET AL., *supra* note 84, §23:03 n.229 and accompanying text.

reliability challenges are few and far between. The method contains multiple discrete steps, from the vaporization of the unknown substance into component ions, the passage of those ions through the stationary phase, expulsion into the detector, the separation of the ions by mass, the recording of the grid of ions on the mass spectrum, and the comparison of that mass spectrum to known chemical substances to determine the overall content.[107] Each of these steps is based on a scientific model of a physical system that is both transparent and explainable. In addition, that reliance on scientific models and our understanding of the complete system enable us to replicate the process using a different machine to test the accuracy of the original machine's output.

To the extent that *Daubert* or other reliability screening require the proponent of the evidence to establish that it is "good science," based on a methodology that is founded in the methods and procedures of science, it is clear that GC/MS passes the test.[108] In those rare cases where the admission is challenged, courts acknowledge this conclusion and thus dismiss the challenge without detailed evidentiary hearings. Judge Lawson, in *United States v. Crockett*,[109] is typical in this regard, in denying a motion to exclude the expert opinion on the purity of drug samples in the police crime lab, when he concluded: "Gas chromatography and infrared spectrometry have long been recognized as valid technological methods for determining the composition of physical substances."[110] The only issue, then, would be to ensure the proper foundation for the test.[111]

3. Gas Chromatography and Mass Spectrometry and the Right to Confront

While reliability challenges are unlikely to result in exclusion of Tier One machine-based evidence, the Confrontation Clause does have more direct effects on admission of chemical testing in court.[112] As discussed above, the Confrontation Clause applies to expert opinion testimony;[113] as a result, admission of a chemical test requires that the analyst be available to testify and be subject to cross-examination.[114] For example, in *Bullcoming*, the majority decided that the testimony of a chemical analyst

107. *Id.* §[4][a].
108. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993).
109. United States v. Crockett, 586 F. Supp. 2d 877 (E.D. Mich. 2008).
110. *Id.* at 886.
111. *Id.* at 880.
112. *See* discussion *infra* Section II.B.
113. *See* discussion *infra* Section II.B.
114. Bullcoming v. New Mexico, 564 U.S. 647, 651–52 (2011).

who performs testing on samples but did not perform the specific test in question is insufficient to satisfy confrontation demands.[115]

For many years, however, the extent to which the Confrontation Clause covered expert testimony like chemical analysis was unclear. In *Williams v. Illinois*,[116] the Supreme Court found that the Confrontation Clause did not require cross-examination about the underlying data used to form an expert's opinion because such data is not "testimonial" and thus not protected by confrontation rights. Several cases interpreted the *Williams* ruling to find that so long as the materials used by the expert are a proper basis of opinion under FRE 703, then they are not testimonial and therefore lack confrontation protection.[117]

The *Smith v. Arizona* ruling in 2024 offers significant clarity on this point, finding that "[i]f an expert for the prosecution conveys an out-of-court statement in support of his opinion, and the statement supports that opinion only if true, then the statement has been offered for the truth of what it asserts."[118] If so, the confrontation right would attach to a chemical test used by a second analyst to conclude that narcotics are present.[119] Remember, however, what the right would entail. The Court in *Smith* made clear that if the statement was offered for the truth, then the accused would have the right to question the test, i.e., to ensure the test was done, the protocols were followed, and the results were accurate.[120] What the Court is demanding, then, is an assessment of whether the result is *valid*.[121] If so, this has implications for AI-based evidence, as discussed below.[122]

The application of confrontation in this type of situation is not so new as to be unprecedented. At least one state supreme court has applied the *Smith* approach for nearly a decade already. The California Supreme Court found, in *People v. Sanchez*,[123] that "case-specific hearsay" would violate the Confrontation Clause if presented without the ability to confront the speaker, since the analyst must be relying upon those facts as the truth to reach a conclusion.[124]

One other note is in order. Some cases have also made a distinction between the assessment of an issue by the analyst and an assessment by the machine itself. If the analysis is truly a product of solely the computer

115. *Id.* at 652.
116. Williams v. Illinois, 567 U.S. 50, 57–58 (2012).
117. *See, e.g.*, United States v. Pablo, 696 F.3d 1280, 1287–88 (10th Cir. 2012).
118. Smith v. Arizona, 602 U.S. 779, 795 (2024).
119. *Id.* at 796–99.
120. *Id.* at 798.
121. *Id.*
122. *See* discussion *infra* Section IV.B.1.
123. People v. Sanchez, 374 P.3d 320 (Cal. 2016).
124. *Id.* at 334–35.

or machine, then some courts have been unwilling to find confrontation applies to these conclusions.[125] A North Carolina GC/MS case, *State v. Pless*, is illustrative of the distinction being made here.[126] In *Pless*, the analyst testified that the substances in question were heroin and oxycodone; however, the forensic chemistry analysis was done by another analyst who was unavailable for trial.[127] The court decided first that the conclusion of the court in a prior case—that confrontation did not apply to the underlying data the analyst used to create her opinion—resulted in the denial of the confrontation challenge on appeal.[128] This initial point will undoubtedly not survive the holding in *Smith*.[129] However, the *Pless* court had a second holding.[130] The court noted that one component of the chemical analysis, the assessment of the weight of the substance, is entirely machine-generated.[131] Therefore, the court decided, it is not a statement at all, and it cannot be either hearsay or testimonial.[132] While the court did not do a deep analysis, the opinion is reminiscent of Justice Sotomayor's concurrence in *Bullcoming*, in which she concludes that truly machine-only analysis does not have confrontation protection since it is from the machine, not a statement of a witness.[133]

4. Standard DNA Analysis with Computer-Assisted Electrophoresis

Before we can discuss our computerized assessment of DNA mixtures, our Tier Two system, we need to first review standard DNA analysis with computer-assisted electrophoresis, the Tier One system upon which it relies. In order to connect a DNA sample from a crime scene to a known DNA sample, a crime lab will follow a standard procedure. First, there will be an amplification of the sample using a Polymerase Chain

---

125. State v. Pless, 822 S.E.2d 725 (N.C. Ct. App. 2018).
126. *Id.*
127. *Id.* at 726–27.
128. *Id.* at 731–32.
129. Smith v. Arizona, 602 U.S. 779, 795 (2024).
130. *Pless*, 822 S.E.2d at 732.
131. *Id.*
132. *Id.* (*citing* State v. Ortiz-Zape, 743 S.E.2d 156, 162 (N.C. 2013)).
133. Bullcoming v. New Mexico, 564 U.S. 647, 673–74 (2011) (Sotomayor, J., concurring in part). More recent caselaw in other contexts supports this contention. In a series of cases regarding electronic data recorders from automobiles, courts have found that the EDR data is the product of the computer itself and, thus, not a statement from a human witness (or the programmers). *See* Nguyen v. State, No. 05-20-00241-CR, 2002 WL 3714494 (Tex. App. 2022); State v. Ziegler, 855 N.W.2d 551 (Minn. Ct. App. 2014).

Reaction (PCR) method,[134] then the analyst will use a genetic analyzer, which sends portions of the DNA through a small transparent tube and, by so doing, creates a graph of the size and intensity of each component.[135] The resulting electropherogram will then contain the allele pattern of the sample being analyzed,[136] and that pattern can finally be compared to other samples, known or unknown, to reach a conclusion on whether the samples match or there is an exclusion.[137] If there is a match, the probative force of the match can be explained in terms of a "random match probability," that is, the likelihood that a random person from the population selected at random would have the same genetic profile at the alleles tested.[138] That information helps clarify the probative force of a DNA match. Each portion of the analysis contains multiple steps in order for the result to be accurate and reliable.

The first step of the analysis, once the DNA is extracted from the sample, is to amplify the targeted regions of DNA using the PCR technique.[139] The PCR method, in general, is designed to take the original sample, isolate the strands at the alleles of interest, and then copy them repeatedly using chemical replication for an exact copy.[140] This is done by taking the overall DNA sample, a double helix of complementary base pairs, and separating it into single strands by heating it to a high temperature.[141] Then the new strand is "clipped" using a chemical primer, which isolates the allele location that is of interest to the analysis.[142] At this point, there will be a single strand of DNA, solely in the area of interest, which can be used to create a complementary strand of base pair segments when heated and combined with additional nucleotide blocks.[143] When complete, the DNA strand will now, once again, be a double helix, and since the same process occurred with the complementary strand separated at the start, there are now two identical sets of the segments of

---

134. FAIGMAN ET AL., *supra* note 84, § 30:44; PAUL GIANNELLI ET AL., *supra* note 84, § 18:03[b]. *See also* David H. Kaye & George Sensabaugh, *Reference Guide on DNA Identification Evidence*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 143 (3rd ed. 2011) (detailing the DNA amplification process).
135. Kaye & Sensabaugh, *supra* note 134, at 144; FAIGMAN ET AL., *supra* note 84, § 30:45; GIANNELLI ET AL., *supra* note 84, §18:03[b].
136. FAIGMAN ET AL., *supra* note 84, § 30:45; GIANNELLI ET AL., *supra* note 84, § 18:03[4][b].
137. Kaye & Sensabaugh, *supra* note 134, at 159–60.
138. *Id.* at 163–67.
139. *Id.* at 143.
140. *Id.*; *see also* FAIGMAN ET AL., *supra* note 84, § 30:44.
141. Kaye & Sensabaugh, *supra* note 134, at 143.
142. *Id.*
143. *Id.*

interest.[144] The process to double the total number of segments can be repeated so that afterward the analyst will have large numbers of copies of the original DNA strand for analysis.

Once the sample has been amplified using the PCR process, the next step is to make a profile using a genetic analyzer, through a process called capillary electrophoresis.[145] At this step in the process, the amplified sample is placed in a tube or tray along with several known samples for weight comparison.[146] The tube is then placed within the genetic analyzer, a transparent capillary tube, and subjected to an electric field—a process known as electrophoresis.[147] Because of the electric field, the components travel through the capillary to be measured by an electronic camera, which records the intensity and time of each component.[148] Because of the electromagnetic forces in play, smaller segments will travel through the tube faster than longer segments, so that the camera will record the smallest segments first and larger ones afterward.[149]

A computer then processes the signal from the camera and produces an electropherogram,[150] which shows the sequence of the fragments passing before it as well as their intensity.[151] This electropherogram can separate out segments that originate at different alleles, as the different alleles are stained using dyes before the electrophoresis begins.[152] Thus, the electropherogram will tell the analyst the number of base pairs at each allele location, and that base pair pattern can be compared to databases of known population samples to determine the known allele pattern at each location and its frequency.[153]

Finally, once there is a known allele pattern of the sample, the sample can be compared to a known sample to determine if the samples match. If the unknown sample is different from the known sample at any allele location, then the source of the known sample can be eliminated as a possible source of the unknown sample.[154] If the two samples match at each allele location, however, then the source of the known sample could also be the source of the unknown sample, although it does not definitively

---

144. *Id.*
145. *Id.* at 144.
146. *Id.*; *see also* FAIGMAN ET AL., *supra* note 84, § 30:45; GIANNELLI ET AL., *supra* note 84, § 18:03[5][b] (detailing "Automated STR Analysis").
147. Kaye & Sensabaugh, *supra* note 134, at 144.
148. *Id.*
149. *Id.*
150. *See* FAIGMAN ET AL., *supra* note 84, § 30:45 n.1.
151. Kaye & Sensabaugh, *supra* note 134, at 144.
152. *Id.* at 145–46.
153. *Id.* at 163–64. *See also* GIANNELLI ET AL., *supra* note 84, § 18:03[5][a].
154. Kaye & Sensabaugh, *supra* note 134, at 159.

prove that it is.[155] Instead, the probative force of the match must be expressed in the form of how likely or unlikely such a match would be, if the true source was not the suspect; this likelihood that someone else taken at random would be the source of the matching sample is referred to as the "random match probability."[156] It can be calculated using a product rule, a mathematical principle that the likelihood of independent events occurring is the product of the probability of each event occurring.[157] Thus, the probability of two coin flips both resulting in "tails" is one in four, as the probability of tails for each flip is one-half and the product rule tells us to multiply the probability of those two events for the collective probability of both.[158] Similarly, the likelihood of any allele pattern in a DNA sample is known within a specific population, and since the United States now uses twenty allele locations in a standard DNA profile,[159] the overall random match probability for the entire sample is the product of each of the twenty individual probabilities. Because of the large number of allele segments to be multiplied, a RMP of one in a trillion or smaller is not uncommon.[160]

5. Standard DNA Analysis with Computer-Assisted Electrophoresis is Reliable

Even with the scientific precision of PCR amplification and capillary electrophoresis, the method is not infallible. Of course, intentional misconduct would negate the validity of any DNA analysis, but it is not the sole source of error.[161] Starting at the beginning of the analyst's work, one problem that might affect the validity of the output of the analysis is that the initial sample is either too small or degraded. Standard DNA analysis protocols suggest that a sample size of one billionth of a gram is sufficient for assessment, but for smaller samples, the process of amplification could possibly introduce errors.[162] Similarly, if the sample is

---

155. *Id.*
156. *Id.* at 164.
157. *Id.* at 165–66.
158. *Id.*
159. GIANNELLI ET AL., *supra* note 84, § 18:03, n.162 and accompanying text (citing *Frequently Asked Questions on CODIS and NDIS*, FED. BUREAU OF INVESTIGATION, https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet [https://perma.cc/3DTG-RXDQ] (last visited Dec. 22, 2025)).
160. GIANNELLI ET AL., *supra* note 84, § 18:03, n.149 and accompanying text.
161. Regarding unintentional lab errors in DNA testing as a general matter, see Kaye & Sensabaugh, *supra* note 134, at 153–59 (discussing lab protocols, documentation, and best practices for specimen handling in DNA testing).
162. *Id.* at 151.

of sufficient size but has been exposed to the elements, it may have degraded to a point where all alleles cannot be assessed.[163]

Even with a sufficient sample size without degradation, the PCR amplification process can also be a source of error. When amplifying a sample, the process depends on the exact replication of each strand of DNA within the sample.[164] The process is, as described above, a multistep process with a complex mixture of chemical primers and free nucleotides, and requires detailed training and attention to detail.[165] The *Reference Guide on DNA Identification Evidence* suggests that contamination from foreign DNA, small strand size, or mutations in the area of a primer can all result in imperfect replication.[166]

Once the profile has been established, the next step is to compare it to the known sample or a database of known samples, to see if there is a match and, if so, what the random match probability would be.[167] Yet not all samples lead to definitive results—a sample might match in some areas, but due to degradation, small sample size, or other anomalies, cannot show either definitive exclusion or a match. Experts may then disagree on the result, even with a properly conducted analysis of the sample through capillary electrophoresis.[168]

Just as with GC/MS, DNA analysis is commonly admissible in court. Having reviewed the methodology of analysis using capillary electrophoresis, it is not hard to see why challenges to this procedure fail, since each step is specific, explainable, and discrete. Moreover, each of the steps in the analysis is based upon a transparent and explainable scientific model. For example, the extraction of DNA from red blood cells in a sample, rather than from white blood cells, is based on a model of human biology in which white blood cells contain DNA while red blood cells do not.[169] Similarly, the amplification of DNA is understood within the context of a scientific model of DNA.[170] In addition, the computer

---

163. Degradation will negatively affect larger samples first, thus making them less likely to be available after significant degradation has occurred. *Id.* at 152–53.
164. *See supra* text accompanying notes 76–80.
165. Kaye & Sensabaugh, *supra* note 134, at 152–56.
166. *Id.* at 143–44.
167. *Id.* at 159–67.
168. *Id.* at 160.
169. Red blood cells have neither a nucleus (the primary source of DNA in a human cell) nor do they have mitochondria (a secondary source of DNA in a human cell). *See, e.g.*, People v. Soto, 981 P.2d 958, 963 n.7 (Cal. 1999).
170. For example, our model of DNA tells us that the backbone of each DNA strand is made up of phosphorous, carbon, and oxygen atoms that form a covalent bond. This allows us to identify enzymes (called "nucleases") that are capable of cutting DNA into pieces at the locations of interest. *See, e.g.*, JEREMY W. DALE & MALCOLM VON SCHANTZ, FROM GENES TO GENOMES 7 (2d ed. 2007).

processing of the image from the electrophoresis relies on well-known statistical models.[171]

When the procedure was first developed, federal courts applied the newly created *Daubert* test to determine whether it met the requirements of admissibility. Thus, in a series of circuit court opinions like *United States v. Beasley*[172] and *United States v. Hicks*,[173] judges performed a multi-factor analysis to determine the reliability of these techniques, finding the science was reliable enough to admit under FRE 702.[174]

Today, commentators like Paul Giannelli and David Faigman describe the state of basic reliability challenges as follows: "Courts now universally accept as generally reliable both the PCR process for amplification of DNA, and the STR-based system of identifying and comparing alleles"[175] and; "As of this writing, STR testing is now widely admissible in virtually all jurisdictions."[176] Once general reliability of DNA testing was established, defense challenges shifted to more specific issues, such as the reliability of random match probability calculations,[177] the use of the product rule,[178] or case-specific lab protocols.[179] Barring any particular error in the case, the admission of DNA evidence using the PCR amplification and capillary electrophoresis methodology is widely recognized.

6. Computer Assisted Electrophoresis and the Confrontation Clause

Confrontation challenges to basic DNA analysis are also unlikely to succeed. The Sixth Amendment guarantees the right to confront witnesses

---

171. *See, e.g.*, ANDREW J. LINK, 2-D PROTEOME ANALYSIS PROTOCOLS 363–81 (discussing computer processing of electrophoresis output); James I Gareels, *Two-Dimensional Gel Electrophoresis and Computer Analysis of Proteins Synthesized by Clonal Cell Lines*, 254 J. BIOLOGICAL CHEMISTRY 7961–77 (1979) (describing an early system for analyzing electrophoresis out by using a computer algorithm).

172. 102 F.3d 1440 (8th Cir. 1996).

173. 103 F.3d 837 (9th Cir. 1996).

174. *See also* United States v. Shea, 957 F. Supp. 331 (D.N.H. 1997); United States v. Lowe, 954 F. Supp. 401 (D. Mass. 1996); *see generally* FAIGMAN ET AL., *supra* note 84, § 30:6; GIANNELLI ET AL., *supra* note 84, § 18:05 (explaining the judicial assessment of reliability of DNA evidence led to widespread acceptance of the procedure as reliable enough to be admissible in court).

175. GIANNELLI ET AL., *supra* note 84, §18:05 n.419 and accompanying text.

176. FAIGMAN ET AL., *supra* note 84, § 30:6.

177. United States v. Gaines, 979 F. Supp. 1429 (S.D. Fla. 1997).

178. *See, e.g.*, People v. Soto, 981 P.2d 958 (Cal. 1999); People v. Reeves, 109 Cal. Rptr. 2d 728 (Cal. Ct. App. 2001).

179. *See, e.g.*, People v. Venegas, 954 P.2d 525 (Cal. 1998); People v. Castro, 146 Misc. 2d 956 (N.Y. Gen. Term 1989). *See generally* GIANNELLI ET AL., *supra* note 84, § 18:05[3].

in a criminal trial, and that includes expert testimony.[180] So, just as with chemical testing, the right to confront witnesses would likely include the ability to ask questions of the analyst who actually performed the analysis in the case, as *Melendez-Diaz*[181] and *Bullcoming*[182] make clear.[183] Similarly, the *Williams* decision sowed confusion when it deemed the use of findings from an outside lab, when used as a basis for a DNA expert's opinion, to be nontestimonial and therefore not subject to the Confrontation Clause.[184]

The courts' willingness to permit an expert to rely on an another person's out-of-court statement in forming an opinion without confrontation protection ended with *Smith* in 2024.[185] In this case, the Court refused to allow the testimony of a second analyst who relied on the testing of a first analyst, since it only would be a basis of opinion if true and, thus, was offered for the truth.[186] The holding would apply equally to DNA testing, like in *Williams*, so that a second analyst's opinion of a "match" would subject the first analyst's assessment to confrontation. Remember, however, what the right would entail. The Court in *Smith* made clear that if the statement was offered for the truth, the accused would have the right to question the test; i.e. to ensure the test was done, the protocols were followed, and the results were accurate.[187] Just as with Gas Chromatography then, the Court is demanding an assessment of whether the DNA result is ***valid***.[188] This would suggest how confrontation would also apply to AI-based evidence, as will be further discussed below.[189]

B. Tier Two Systems: Proprietary and Complex Computer-Aided Analysis

Tier One machine-assisted chemical analysis, like gas chromatography and mass spectroscopy or DNA electrophoresis, are admissible precisely because each step in the process meets the requirements of transparency and explainability.[190] Tier Two systems are

180. *See* discussion *supra* Section II.B.
181. Melendez-Diaz v. Massachusetts, 557 U.S. 305, 324 (2009).
182. Bullcoming v. New Mexico, 564 U.S. 647, 652 (2011).
183. *See supra* text accompanying notes 41–43.
184. *See* Williams v. Illinois, 567 U.S. at 57–58.
185. Smith v. Arizona, 602 U.S. 779 (2024).
186. *Id.* at 795.
187. *Id.* at 798.
188. *Id.* at 799.
189. *See* Section V.B.1.
190. See discussion *supra* Sections IV.A.2 and IV.A.5.

similar to Tier One systems but also include complex computer algorithms as part of the system. Because a computer algorithm is simply another step in the overall process, Tier Two systems are not intrinsically different from Tier One systems. The difference arises when those computer systems are proprietary or complex. To understand the issues facing the admissibility of Tier Two machine-based evidence, it is necessary to review the mechanism by which machines make comparisons using the example of computer assisted DNA mixture analysis.

1. Computerized Assessment of DNA Mixtures

Standard DNA analysis using Tier One methods is a transparent process of detailed assessment, which can be assisted by computers but lacks the complexity to raise significant reliability or confrontation concerns under current doctrine. Yet, in a series of cases involving a specific subset of DNA analyses, the science and the courts have been pressed to the absolute limit of current doctrine. These cases involve computerized assessment of mixtures.

A DNA mixture is a sample that contains fluid or tissues from more than one individual.[191] In a standard electropherogram, each locus will contain two allele segments for each individual—one inherited from the father and one inherited from the mother. In a mixture, the number of allele segments per locus will be above two since each locus will contain two segments from each contributor. Thus, if an analyst sees more than two segments at different loci, the analyst can conclude that the sample is a mixture.[192]

The simplest case of identifying a genetic profile from a mixture is when there are two contributors to a sample and one is known (e.g., the victim).[193] If this is the case, the lab can obtain an individual profile from the known contributor, and then "subtract" the known allele patterns from the victim from the electropherogram results, leaving the profile of the other contributor to the mixture remaining.[194] A different, but also relatively simple, response can be used when the crime victim and suspect are of the opposite sex. Since men have one X chromosome and one Y chromosome, the male contributor can be identified through the Y chromosome in the mixture.[195] If it is suspected that more than one male is a contributor to the mixture, the analyst can count the number of

191. Kaye & Sensabaugh, *supra* note 134, at 182.
192. *Id.* at 183.
193. FAIGMAN ET AL., *supra* note 84, § 30:59 n.9 text following note 9.
194. Kaye & Sensabaugh, supra note 134, at 184–85.
195. FAIGMAN ET AL., *supra* note 84, § 30:59 nn.8–9 at text accompanying notes 8–9.

different Y chromosomes at that location, in addition to determining whether the suspect has a Y chromosome consistent with one of those present in the sample.[196]

More complex problems arise, however, when the number of contributors to the mixture is unknown. In these circumstances, the analyst can make an informed guess at how many contributors exist and then, based on the allele patterns, try to deduce which alleles belong to which contributor.[197] Making these choices about allele patterns involves a "good deal of judgment," and, in recent years, has involved the application of statistical assessment of computer systems to "deconvolute" the mixture and determine the most likely number of contributors.[198] This software, and the presentation of the results from its analysis, has led to a series of court cases that have significant importance to the assessment of AI.

To "deconvolute" the mixtures, computer systems will use statistical modeling to assess the likelihood of different combinations, which Faigman describes as "too difficult to 'deconvolute' by hand."[199] These systems are highly dependent upon the use of a probabilistic model that typically provides a likelihood that the particular set of DNA fragments in the mixed-DNA sample are from an individual suspect.[200] Such likelihood ratio can be described as:

$$LR = \frac{P(O = o|H_1, G_s = g_s)}{P(O = o|H_2, G_s = g_s)}$$

Where LR means likelihood ratio;

P($x|a$, b) is the probability of x given that a and b are true;

O is the epg[201] of the crime stain profile of the sample;

o is the epg of the crime stain profile of the individual suspect;

196. Kaye & Sensabaugh, *supra* note 134, at 184.
197. FAIGMAN ET AL., *supra* note 84, §30:59 n.10 at text accompanying note 10
198. Kaye & Sensabaugh, *supra* note 134, at 185; FAIGMAN ET AL., *supra* note 84, §30:59 nn.10–11 at text accompanying notes 10-11.
199. FAIGMAN ET AL., *supra* note 84, § 30:34.
200. *See, e.g.*, Maarten Kruijver et al., *Exploring the Probative Value of Mixed DNA Profiles*, 41 FORENSIC SCI. INT'L: GENETICS 1, 3 (2019).
201. An epg (electropherogram) is a graphical representation of the length a base pair. *See* R. G. Cowell et al., *Analysis of DNA Mixtures With Artefacts*, 64 J. ROYAL STAT. SOC'Y SERIES C: APPLIED STATS. 1, 5–6 (2015).

H_1 is the hypothesis that the individual suspect is a contributor to the mixture;

H_2 is the hypothesis that the individual suspect is not a contributor to the mixture;

G_s is the genotype of the sample; and

g_s is the genotype of the individual suspect.[202]

Due to the complexity of this mathematical analysis and "[k]nown shortcomings of traditional methods of DNA profile interpretation," the use of AI systems (e.g., expert systems) has been brought to help with the analysis.[203] Such expert systems add a further layer of modeling. For example, the STRmix™ system incorporates a model of "drop-in [DNA] (the presence of law amounts of extraneous DNA) and dropout [DNA] (which is a consequence of law template and/or degraded DNA and results in partial DNA profiles)."[204]

These computer systems will be tested repeatedly before they are used in real cases, so they have passed an initial "developmental validation" phase before they are sent to labs for use, who will also perform separate validation.[205] The validation might lead to the conclusion the systems are reliable, but a recent report from the National Institute for Standards and Technology, addressing this issue, instead concluded that "[c]urrently, there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices, including the use of probabilistic genotyping software (PGS) systems."[206]

2. DNA Mixture Analysis and Reliability Concerns

Some courts have been skeptical of the reliability of computer assessment of mixtures, leading to decisions that exclude the results of

---

202. *See id.* Although, "[w]hen the [sample] originates from a single individual, the weight of evidence can be presented as a match probability." Jo-Anne Bright et al., *Developmental Validation of STRMix™, Expert Software for the Interpretation of Forensic DNA Profiles*, 23 FORENSIC SCI. INT'L: GENETICS 226, 226 (2016).
203. *See* Bright et al., *supra* note 202, at 227.
204. *See id.* at 226–27.
205. *Id.*
206. FAIGMAN ET AL., *supra* note 84, §29:20 (citing NAT'L INST. SCI. & TECH., DNA MIXTURE INTERPRETATION: A NIST SCIENTIFIC FOUNDATION REVIEW 6 (2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8351-draft.pdf [https://perma.cc/KHL2-8ENN].

those analyses. In a 2019 opinion, *Morten v. State*,[207] the Maryland Court of Special Appeals held that while the results from TrueAllele (a commercially available program) were reliable enough to admit, the defense, nonetheless, retained the right to present evidence to the jury challenging that reliability.[208] That same year, a federal district court in Michigan found, in *United States v. Gissantaner*, that a different mixture analysis software—STRMix—was not even reliable enough to admit as a matter of gatekeeping under the *Daubert* standard.[209] In that opinion, Judge Neff found that the software lacked testing, peer review, and general acceptance for low-level mixtures,[210] and that controlling standards were not set for the use of STRMix generally.[211] For those reasons, "the sum of the parts simply [did] not add up to a reliable whole, the . . . use of STRMix probabilistic genotyping software must be excluded."[212] Finally, in 2021, the New Jersey Superior Court found in *State v. Pickett* that the reliability assessment of the mixture software—here, TrueAllele—required the discovery and assessment of the software source code.[213] Specifically, the court noted that TrueAllele had never been subjected to independent analysis by outsiders to determine whether it uses valid scientific methods, but also implements those methods properly within the program.[214]

Yet, these opinions are not the sole voice on the issue. In fact, Judge Neff's opinion is not even the sole voice on the issue in *United States v. Gissantaner*. On appeal to the Court of Appeals for the Sixth Circuit in 2021, the panel overturned Judge Neff's District Court opinion, finding that STRMix need not be test*ed* to be test*able*, that peer review need not be from independent reviewers, and that it has been admitted by other courts.[215] Two decisions since the *Gissantaner* appellate decision reach similar results. In 2022, the New York Court of Appeals, in *People v. Wakefield*, found that the TrueAllele program was reliable enough to admit

207. Morten v. State, 215 A.3d 846, 873–75 (Md. Ct. Spec. App. 2019).
208. *Id.*
209. United States v. Gissantaner, 417 F. Supp. 3d 857 (W.D. Mich. 2019), *rev'd*, 990 F.3d 457 (6th Cir. 2021). Regarding the handling of this case on appeal to the Sixth Circuit see *infra* text accompanying note 200.
210. In the *Gissantaner* case, the DNA sample was described as "a complex mixture of low-template touch DNA consisting of at least three contributors in which the person of interest is determined to be a minor contributor of only 7%." *Id.* at 875.
211. *Id.* at 879 (testing), 882 (standards, for low-level mixtures and STRMix generally), and 884 (general acceptance).
212. *Id.* at 887.
213. State v. Pickett, 246 A.3d 279, 310 (N.J. Super. Ct. App. Div. 2021).
214. *Id.* at 300–01.
215. *Gissantaner*, 990 F.3d at 463–64 (emphasis added).

at trial.[216] Specifically, the court found that TrueAllele's methodology for mixture analysis was "better than a human analyst's assessment of the same issue.[217] The court also rejected the notion from cases like *Pickett* that source code would be necessary for independent assessment of the software.[218] In 2023, Judge Brann of the Middle District of Pennsylvania also found TrueAllele reliable enough to admit after a *Daubert* hearing and based on the *Wakefield* and *Gissantaner* opinions among other *Daubert* considerations.[219]

In sum, while reliability assessment of the standard procedures of DNA analysis is rare, the issue of reliability for mixture analysis, specifically computer-aided statistical modeling software for it, remains an open question. But, the handling of the issue by courts like *Gissantaner*, *Wakefield*, *Morten*, and *Pickett* provide valuable insight into the handling of similar challenges for AI-based expert evidence.

3. Beyond Reliability: Transparency and Explainability

As demonstrated above, courts have split on the reliability of computer-aided DNA mixture analysis.[220] Some, like the Court of Special Appeals of Maryland in *Morten v. State* and the District Court for the Western District of Michigan in *United States v. Gissantaner*, have raised concerns about the reliability of such systems.[221] Others, like the Sixth Circuit in *United States v. Gissantaner* and the New York Court of Appeals in *People v. Wakefield*, did not see any real reliability issues.[222]

Unfortunately, all of these opinions fail to see the true problem with proprietary algorithms.[223] It is impossible to determine whether a complex

216. People v. Wakefield, 195 N.E.3d 19, 21 (N.Y. 2022) (finding that the TrueAllele evidence was reliable enough to admit under New York's *Frye*-based general acceptance standard).

217. *Id.* at 28–29 (finding the empirical data in validation studies of TrueAllele shows that the methodology is generally accepted and that the source code need not be disclosed, due to the validation studies).

218. *Id.*

219. United States v. Anderson, 673 F. Supp. 3d 671, 688 (M.D. Pa. 2023). Six months later, Judge Jackson of the Middle District of Louisiana relied heavily on Neff's opinion in *Anderson* to admit TrueAllele results in *United States v. Lockett*, No. 20-00091-BAJ-RLB, 2023 WL 7181251 (M.D. La. Nov. 1, 2023).

220. *See supra* text accompanying notes 188–200.

221. Morten v. State, 215 A.3d 846, 873–75 (Md. Ct. Spec. App. 2019); United States v. Gissantaner, 417 F. Supp. 3d 857, 879 (W.D. Mich. 2019).

222. United States v. Gissantaner, 990 F.3d 457, 463 (6th Cir. 2021); *Wakefield*, 195 N.E.3d at 28.

223. *See Morten*, 215 A.3d at 869–70 (discussing how it was improper to limit expert's testimony as to theoretical and ad hoc reliability of TrueAllele testing); *Gissantaner*, 417

computer algorithm will be reliable in a given case without access to the code (transparency) and the ability to understand it (explainability).

We can understand this problem by imagining various responses to questions asked of a scientific expert to assess whether the expert has used a reliable machine-based system. Let us start by asking our hypothetical expert to explain how they determined that the sample was methamphetamine. Imagine that the expert answered, "science magic." Certainly, that would not meet the standard for admissibility because the judge would have no ability to identify the scientific model or principle involved For that reason, the judge would not be able to assess the methodology's reliability.[224] Now further imagine that when pressed on how they knew their science magic worked, they said, "we ran a lot of tests, and it correctly identified the substance 98% of the time." Here the expert is attempting to justify using their magical system by validating that system—using statistics to argue that it is accurate.[225] Is validation sufficient in such a case?

We contend that it is not. Scientific evidence must be derived using a scientific model. Simply stating that your model has been validated provides no ability to assess whether the model used is scientific. For example, the *Daubert* court identified four (non-exclusive) factors to assess whether evidence was scientific: testability, known error rate, peer review, and general acceptance.[226] If all we know is that a complex algorithm has been validated, we can only assess one of these factors. As discussed below, showing validation is insufficient to establish

F. Supp. 3d at 879 (focusing on questions about testing and validation of the software in comparable situations as that presented at trial); *Gissantaner*, 990 F.3d at 463 (applying the *Daubert* factors to assess reliability); *Wakefield*, 195 N.E.3d at 30 (rejecting defendant's arguments about accessing the source code because the arguments related to how accurately the code performed and if it did so as intended).

224. *See* United States v. Frazier, 387 F.3d 1244, 1265–66 (11th Cir. 2004) (finding that the trial judge did not abuse their discretion by excluding expert witness testimony for which the witness failed to explain the basis for their opinion); Rodrigues v. Baxter Healthcare Corp., 567 Fed. App'x. 359, 361–62 (6th Cir. 2014) (finding that the trial court did not abuse its discretion by excluding testimony of expert witness who could not explain the process underlying testimony); Johnson v. J.P. Parking, Inc., 717 F. Supp. 3d 798, 815–17 (S.D. Iowa 2024) (excluding testimony of an expert witness who did not explain the reason why he chose certain values as components for a damages calculation).

225. Validation generally relies on the developer of the instrument "acqui[ring] test data to verify the functionality of the system, the accuracy of statistical calculations and other results, the appropriateness of analytical and statistical parameters, and the determination of limits." *See, e.g.*, SCI. WORKING GRP. ON DNA ANALYSIS METHODS, GUIDELINES FOR THE VALIDATION OF PROBABILISTIC GENOTYPING SYSTEMS 5 (2015), https://www.swgdam.org/_files/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf [https://perma.cc/32HF-B2TA].

226. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593–94 (1993).

reliability.[227] Thus, we need to know the model upon which the algorithm is based—it must be transparent (we know what it is) and explainable (understandable to the parties).

Let us take our example one step further. Imagine that when pressed the expert states that the science magic used was differential migration. At this point the judge is able to assess the *theoretical reliability* of the system.[228] In essence, the judge can determine if differential migration is based on a scientific model or principle. The judge then asks for information about the apparatus used to perform the differential migration. Here, the expert demurs arguing that the apparatus is proprietary, so they cannot provide the details; but they also reassure the judge by stating that the system was validated and is based on a scientific model.

At this point, the judge may want to accept the evidence as based on a scientific model (differential migration) that is reliable (validated). But that is an error. Anyone can claim they have built an apparatus using a scientific model and validated it, but that does not by itself tell us that the system is reliable. As explained below, validation does not establish that the system as developed or implemented is reliable in a new circumstance (called "ad hoc" reliability[229]).

The validation argument misses a fundamental point. A statistically reliable system can produce statistically reliable results based on a flawed or even irrelevant model. The case of Clever Hans provides an interesting, historical example of this type of problem. Starting in 1891, a horse by the name of Clever Hans became a scientific sensation when it was found that he "could apparently perform mathematical calculations, tell time, identify musical intervals, and name people."[230] Relative to mathematical calculations, Hans was able to correctly answer about 90% of the time by tapping out the answer.[231] Given such a high rate of correct answers, whatever Clever Hans's methodology was, it appeared very reliable. Sixteen years after the world was introduced to Clever Hans, "a group of thirteen scientists (the 'Hans Commission') re-tested Clever Hans" using a very carefully controlled psychological experiment.[232] As a result, the scientists learned that Clever Hans could not actually do math, rather, he

227. *See infra* text accompanying notes 211–19, 205–15.

228. An attack on theoretical reliability of a scientific model is an attack on the "scientific" nature of the model. *See* Morten v. State, 215 A.3d at 869.

229. An attack on ad hoc reliability would be an argument that relates to the reliability of the manner in which the system was built, used, or used. *See id.* at 870.

230. Tegoseak v. State, 221 P.3d 345, 351 n.7 (2009).

231. Wojciech Samech & Kaus-Robert Müller, *Towards Explainable Artificial Intelligence*, *in* EXPLAINABLE AI: INTERPRETING, EXPLAINING AND VISUALIZING DEEP LEARNING 7 (Wojciech Same et al. eds., 2019).

232. *Id.*

was extremely good at reading the questioner's subtle body language to know when to stop tapping.[233] Thus Clever Hans's *reliable* "model" of mathematics had nothing to do with numbers; instead, it was a model of human psychology. And without knowing what that model was, we could never be sure whether Hans would be reliable in a given circumstance.

Similar problems arise in the area of computer algorithms (AI systems). For example, a neural network was trained to distinguish photographs of wolves from husky dogs.[234] Once trained, the system was highly reliable.[235] But this reliability was a mirage as the investigators eventually realized that the system was able to distinguish wolves from huskies not based on the difference between the animals, but because all of the pictures of huskies, in the development and test data, were in snow—so the system identified the husky as a husky (and not a wolf) based on a model of snow.[236] As one can imagine, such a system's reliability would radically change if shown pictures of huskies living in, for example, Florida. But until you know what the model is, you cannot know in what circumstances it would be unreliable.

The central point here is that validation can only take us so far. Due to the complexity of creating these computer algorithms, the model we think the system uses may not be what it actually uses. Without access to the model as instantiated in the computer code, we cannot determine if system would provide a correct output in our case.

As a result, all validation shows that the system has produced correct output a certain percentage of the time using the validation data. It does not tell us that the system will be reliable relative to our case because we cannot know if the model used by the system is flawed in a way that gives unreliable results in our circumstances. We can only conclude that the model will produce reliable results in our case if we know what model the systems uses. Such knowledge requires that the model (the code and associated data) be transparent (provided to us) and explainable.

In this context, if the STRMix and TrueAllele systems are to be judged reliable, the relevant code and data associated with the systems must be provided to the court. When proprietary systems are not so provided to the court, then they must not be admitted because the court cannot judge them to be reliable.

---

233. *Id.*
234. Marco Tulio Ribeiro et al., *Why Should I Trust You? Explaining the Predictions of Any Classifier*, *in* PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1135, 1142–43 (2016).
235. *Id.*
236. *Id.* at 1142.

C. Tier Three Systems: Closed Box Neural Networks

Tier Two systems rely on proprietary, complex computer algorithms (models) that, in principle, can be provided to the court and opposing party for analysis and review. As such, their admissibility poses no special issues other than providing access to the proprietary systems—something the courts are well-suited to do.[237] Tier Three systems are those systems in which the model used by the system is *in principle* not explainable even if the computer code is provided to the other side. Such systems are not admissible because they cannot be assessed for reliability.

One important version of Tier Three systems are closed-box neural networks—complex computer systems that cannot be explained because the model they use cannot be deciphered from the system's code.[238] And, while such systems tend to be used in law for non-evidentiary purposes (drafting assistance, document analysis, text retrieval, case/legislation analysis, speech-to-text, chatbots, and office administration),[239] it is likely, given the advances in LLMs like OpenAI's ChatGPT (a closed-box system),[240] that we will see them being used in machine-based evidence in the very near future. As we argue below, machine-based evidence based on such closed-box systems should not be admitted because it is impossible to assess their reliability, and their admission would violate criminal defendants' due process rights.[241]

In order to understand the complexity of neural networks and why they form closed-box systems, we need to explain how a neural network works. A neural network is based on our understanding of the functioning of the brain.[242] The brain is considered to have two low-level features: the neuron (which is the processing unit) and the synapses which connect the neurons (the parameters).[243] Unlike traditional computer systems where instructions are executed serially (one by one), neurons act in parallel—

---

237. *See, e.g.*, Garcia v. Peeples, 734 S.W.3d 343, 348 (Tex. 1987) (discussing protecting proprietary interests in discovery).
238. *See, e.g.*, Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUTING SURVEYS 93:1 (Aug. 2018).
239. *See, e.g.*, HOMOKI, *supra* note 78, at 20–37.
240. *See* Kosinski, *supra* note 79 (noting that ChatGPT is a black-box system).
241. See discussion *infra* Sections IV.C and V.B.2.
242. GUPTA & NAGPAL, *supra* note 73, at 6.
243. *See* Brian Seamus Haney, *Applied Artificial Intelligence in Modern Warfare and National Security*, 11 HASTINGS SCI. & TECH. L.J. 61, 66 (2020).

each neuron can process its information at the same time as any other neuron.[244]

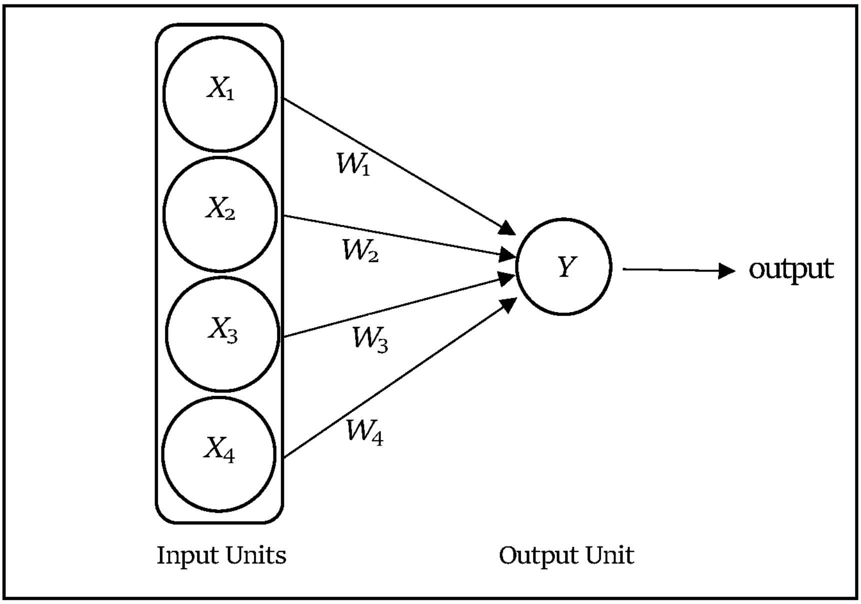

Figure 1: Single Neuron Perceptron[245]

In Figure 1, each X_i is an input unit which has a specified value. These values are transmitted to the output unit (in this case *Y*) by multiplying the value of the input unit by a weight W_i. Y then uses some formula (for example, a weighted sum) to generate the output value.[246] The weights W_1 to W_i are set to some initial value α_1 to α_i and are then modified through analysis of the data (the learning process).[247]

244. ALPAYDIN, *supra* note 61, at 274–75.
245. *Id.* at 275–78.
246. *Id.*
247. *See, e.g.*, *id.* at 278–82.

Most neural networks are not made up of a single perceptron, but rather consist of a combination of multiple perceptrons:[248]

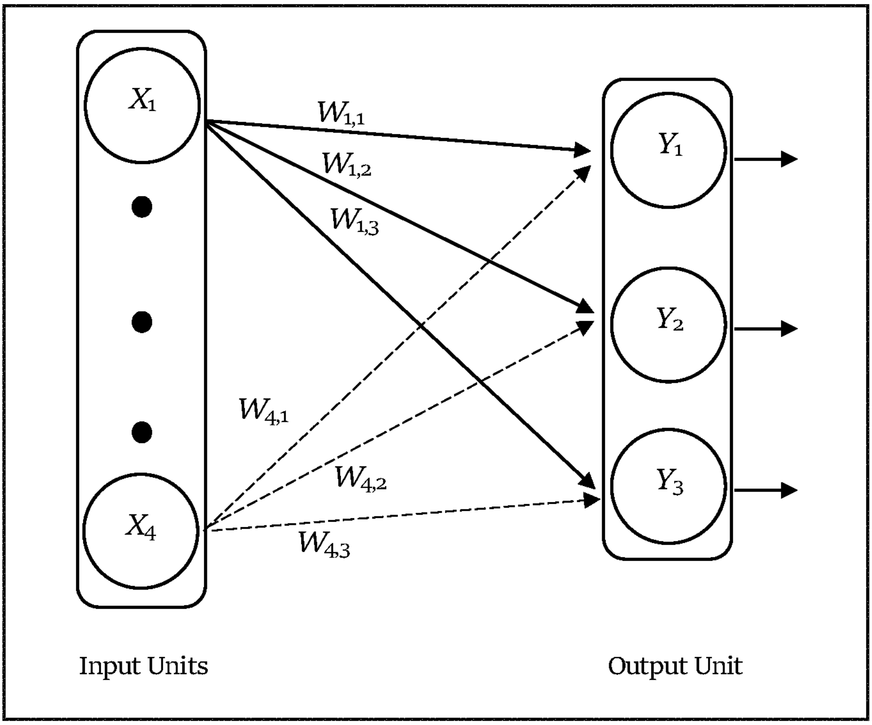

Figure 2: Multiple Parallel Perceptrons

In Figure 2, we still have four input units ($X_1 \ldots X_4$). But now each input unit connects to each of three output units (Y_1, Y_2, Y_3) and we associate a vector of weights $\overrightarrow{W_i}$ (e.g., [$W_{1,1}$; $W_{1,2}$; $W_{1,3}$]) with each input unit.

In the network we have created thus far, there is a direct connection between the input and the output. This implies that there is a linear relationship between the input and the output—a one-to-one correspondence.[249] But sometimes that is not the case. To represent more complex and non-linear relationships, our perceptron model may have multiple (hidden) layers.[250] A network with many hidden layers is called a deep neural network, where "[s]uccessive hidden layers correspond to

248. *See id.* at 283–86.
249. *See id.* at 283.
250. *See* ALPAYDIN, *supra* note 61, at 283–86.

more abstract representations until we get to the final [output] layer."[251] For example, a deep learning network attempting to identify objects in pictorial form might have a hidden layer for edges,[252] another layer might be for corners,[253] and yet another to identify textures.[254] These levels feed into each other until they end in the output—what the picture represents. For example, we could have the following setup:

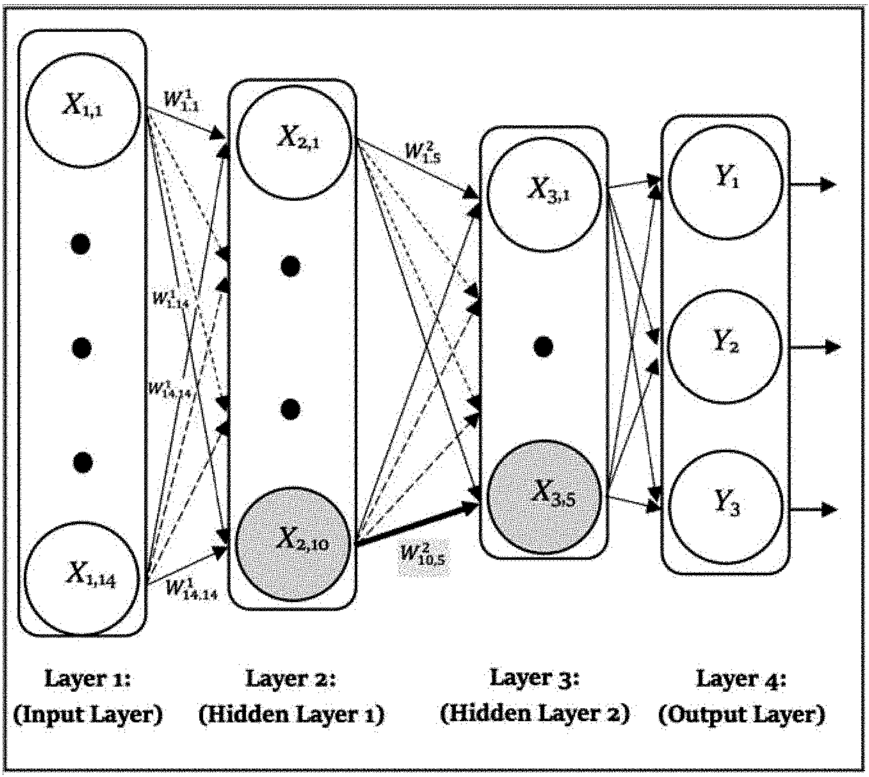

Figure 3: Multiple Parallel Perceptrons

251. *See id.* at 314.
252. *See, e.g.*, Hongli Zhang et al., *A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing*, 7 IEEE ACCESS 159081, 159084 (2019).
253. *See, e.g.*, Jaruwan Toontham & Chaiyapon Thongchaisuratkrul, *The Comparison of Object Recognition and Identification by Using Image Processing Base on the Neural Network, the Hough transform and the Harris Corner Detection*, 1 INT'L J. MACH. LEARNING & COMPUTING 400, 403–04 (2011).
254. *See, e.g.*, R. Kapela, *Texture Recognition System Based on the Deep Neural Network*, 68 BULL. POL. ACAD. SCIS.: TECH. SCIS. 1503 (2020).

Weight $W^{a}_{b,c}$ is the weight of the connection from unit b of layer a to unit c of layer a+1. For example, $W^{2}_{10,5}$ is the weight from layer 2 unit 10 to layer 3 unit 5 (highlighted in Figure 3).

Due to the complexity of the system and the way in which a neural network learns, while the developer may know the logical structure of the neural network and the weights connecting them, the developer may have no idea what model the system is using.[255] As a result, when an expert uses such a closed-box neural network to analyze data, the expert cannot explain why the neural network produces the output it does. This is analogous to the "Clever Hans" and Wolf/Husky problems discussed previously.[256] And, as we have noted, if the expert cannot explain the model the expert relies on to the judge, the judge cannot assess that model's reliability and should not allow such evidence to be admitted.[257]

V. LEGAL CONSTRAINTS ON ADMISSION OF EXPERT TESTIMONY BASED ON MACHINES—RELIABILITY AND CONSTITUTIONAL CONSIDERATIONS

Admission of expert testimony relying on machine-based evidence is subject to the judge's assessment that the evidence is reliable and, in criminal cases, to constitutional limits. The reliability assessment depends entirely on the type (tier) of machine creating the output.[258] In addition, when evidence is being offered against a criminal defendant, constitutional limits of confrontation and due process also shape admissibility.[259] Perhaps surprisingly, the more important of these considerations, we argue, will be the constraints of due process, as confrontation considerations are largely subsumed within the evidentiary analysis of reliability.[260] For that reason, we offer a new constraint on the admission of expert testimony relying on machine-based evidence: that due process requires that the defendant have access to the system's architecture, including, in the case of computer-based evidence, the code and associated data, allowing assessment by trained experts provided (where necessary) to the accused to ensure proper screening of evidence prior to admission.

255. *See* GUPTA & NAGPAL, *supra* note 73, at 265.
256. *See supra* text accompanying notes 232–38.
257. *See supra* text accompanying notes 1–4.
258. *See infra* text accompanying notes 275–82.
259. *See* discussion *infra* Section V.B.
260. *See* discussion *infra* Section V.B.2.

A. Reliability: A Consideration of the Model Used for Analysis

Having reviewed the models for machine analysis, the application of judicial gatekeeping standards to them for reliability screening becomes an easier task. The reliability screening for expert testimony based on machine-based evidence, like with any type of scientific evidence, is focused on the principles and methodologies (the process) by which the system creates its final conclusion.[261] For Tier One and Tier Three systems, reliability analysis of the methodology of decision making leads to a straightforward result of acceptance (Tier One)[262] or rejection (Tier Three).[263] The mixed methods model (Tier Two), however, is more complex and requires additional safeguards before reliability is established.

First, imagine a typical Tier One analysis such as gas chromatography. Having reviewed the methodology by which a gas chromatograph/mass spectrometer evaluates the chemical composition of an unknown substance in detail, it becomes clear why this type of analysis is both transparent and reliable.[264] The methodology is transparent because the process for identifying a substance is clear and well-known. The sample is vaporized into the stationary phase leading to a reading on the detector.[265] Once read, the spectrum of the substance is then compared to signatures of other known substances leading to an identification.[266] Thus, the method by which an identification is made from these steps is clear to the operator and, if need be, can be detailed in testimony in court. Independent of the transparency of the method, it can also be explained since it relies on basic scientific models of the chemical properties of elements, compounds, and ions. As a result, the test can be broken down into principles for each phase of the analysis, and the analysis can be replicated with the same or with a different machine. The combination of the transparency of the process and the explainability of the steps of analysis lead to the process being deemed

261. *See* Daubert v. Merrell Dow Pharms., 509 U.S. 579, 595 (1993) (holding that "[t]he focus, of course, must be solely on principles and methodology, not on the conclusions that they generate").
262. *See* discussion *supra* Sections IV.A.2, IV.A.5.
263. *See* discussion *supra* Section IV.C.
264. For a detailed review of the mechanism of operation of a gas chromatograph/mass spectrometer, see *supra* Section IV.A.1.
265. *See supra* text accompanying notes 86–92.
266. *See supra* text accompanying notes 93–100.

"reliable" under *Daubert*, and thus universally recognized as admissible in court.[267]

DNA analysis using PCR amplification and capillary electrophoresis is similarly transparent and explainable. The process by which an unknown sample is amplified using the PCR technique, then subjected to an electric field leading to an electropherogram, is known, clear, and can be detailed in testimony in court.[268] Independent of the transparency of the methods used to make a comparison, they also rely on known models of the properties of DNA, statistical assessment of allele patterns, and population estimates made from large samples.[269] Just as with gas chromatography, the method can be broken down into principles for each phase of the analysis, and the analysis can be replicated with the same or with a different machine.[270] For similar reasons, courts have universally held standard DNA analysis to be reliable enough to be admissible in court.[271]

These examples are illustrative of machine assisted testimony under a known model (Tier One), and since the methodology of decision-making is both transparent and explainable, the judge can appropriately screen these processes and can find the system reliable under *Daubert*.

Compare Tier One known models to a typical Tier Three closed-box models such as ChatGPT. Programmers will start with a system design with weights assigned to certain inputs, but once the initial design is created it then undergoes a learning process to refine the initial weights leading to new and more complex system architecture.[272] Thus, even if the initial system design was human-created, the ultimate architecture of decision-making will, through the learning process, become opaque through multiple successive layers of perceptrons.[273] The initial developer has now, through the learning process, created a system that is by its very architecture, not transparent.[274] These systems are not transparent because their structure prevents inspection of the methodology of decision-making due to tiered perceptron layers leading to a final result.

267. *Cf.* United States v. Crockett, 586 F. Supp. 2d 877, 886 (E.D. Mich. 2008) (stating "[g]as chromatography and infrared spectrometry have long been recognized as valid technological methods for determining the composition of physical substances").

268. *See supra* text accompanying notes 134–44.

269. *See supra* text accompanying notes 146–60.

270. *See supra* text accompanying notes 134–60.

271. GIANNELLI ET AL., *supra* note 29, § 18:05, n.419 text accompanying note 419. *See also* FAIGMAN ET AL., *supra* note 84, § 30:6, n.6 text following note 6.

272. *See supra* text accompanying notes 248–58.

273. *See id.*

274. *See* GUPTA & NAGPAL, *supra* note 73, at 265.

Nor are these systems explainable, as a final result cannot be traced back to the original source through an identifiable methodology. Even the original designer would not be able to trace the methodology of the decision making process from start to finish, because once the system has been trained using the learning process, it no longer represents the original design of the developer but has changed to become something new and opaque.[275] Furthermore, given the complexity of the system, it is not possible to understand or explain *why* they system produces the output that it does.[276]

The *Daubert* decision made clear that it is the methodology of the analysis, not just the result, that must be reliable.[277] If so, machine models like ChatGPT or similar closed-box neural networks fail even the most basic reliability screening, and are inalterably unreliable and inadmissible in court.

Finally, consider a Tier Two Mixed Model analysis like DNA mixture analysis based on statistical analysis. Unlike standard DNA analysis, a Tier One Known Model, mixture analysis involves complex statistical modeling to deconvolute mixtures deemed too complex to assess manually.[278] The two commercially successful systems for mixture analysis, TrueAllele and STRMix, maintain their code is proprietary information—making the model used by these Tier Two systems unavailable for examination and review.[279] Under these circumstances, the methodology of decision-making is opaque (neither transparent nor explainable) to outside observers and, thus, is distinct from the known procedures of standard (Tier One) DNA analysis.

Yet a significant number of courts have admitted the output of these programs in mixture analysis cases. In facing reliability challenges for

---

275. *See supra* text accompanying notes 248–58.

276.*See* GUPTA & NAGPAL, *supra* note 73, at 265.

277. Daubert v. Merrell Dow Pharms., 509 U.S. 579, 594–95 (1993) (stating that gatekeeping screening will focus on "the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission. The focus, of course, must be ***solely on principles and methodology, not on the conclusions*** that they generate") (emphasis added).

278. *See supra* text accompanying notes 141–43.

279. *See, e.g.*, People v. Wakefield, 195 N.E.3d 19, 35 (N.Y. 2022) (Rivera, J., concurring) (noting that the developer of the TrueAllele program testified that its "the source code was a '***trade secret***,' but that the basic mathematical equations underlying TrueAllele were published in the literature. However, "[t]he engineering of the elaboration of [the equations] in more levels of hierarchy" were ***proprietary***") (emphasis added); State v. Pickett, 246 A.3d 279, 302 (N.J. Super. Ct. App. Div. 2021) (detailing the proprietary nature of TrueAllele before ordering disclosure of source code; noting also the proprietary nature of STRMix, which had been ordered to be disclosed and found to contain code errors).

mixture analyses using TrueAllele, decisions like *Wakefield* in New York and *Anderson* in the U.S. District Court have found the methodology of TrueAllele to be reliable enough to admit.[280] Without the transparency of the source code, however, the rationale for these decisions is based on a combination of output validation and weak reliability screening.[281] Software developers do expend significant efforts in validation of the output of the system before implementation, so at a superficial level this validation could tell the operator and court that the machine is running correctly.[282] But we believe that this approach is fatally flawed, and thus strongly disagree with the approach taken in *Wakefield* or *Anderson*.

The problem with output validation is that it fails to answer the central question that *Daubert* asked judges to determine, that is, whether the methodology of the decision is reliable enough to admit.[283] Justice Blackmon cautioned judges, in performing FRE 702 gatekeeping after *Daubert*, that the analysis could not merely focus on conclusions of the expert, but instead the methodology and principles underlying that conclusion.[284] This bedrock principle cannot be satisfied when the validation of the machine is only through output certification because, having not been given access to the machine or code, you cannot *know* the methodology or principles the machine or code used.[285] All validation establishes is that the model has a specific success rate given the data presented to it. Because you do not know *why* the system produced a given result, you cannot determine what factors might produce unreliable outcomes. In the case of Tier Two systems, like DNA mixture analysis, the proprietary (or complex) nature of the system prevents the process from being transparent or explainable to outside observers. Such a system could, inadvertently, incorporate blind spots, flaws in logic, or other errors

280. *See supra* text accompanying notes 217–20.

281. United States v. Anderson, 673 F. Supp. 3d 671, 682–84 (M.D. Pa. 2023) (relying on validation studies done internally to find TrueAllele reliable under *Daubert*); *Wakefield*, 195 N.E.3d at 30 (relying on validation studies to find reliability under New York's *Frye*-based reliability standard).

282. *See supra* text accompanying notes 141–43. We are not alone in expressing concern about internal validation, in that the NIST DNA Mixture report also recommends independent review. *See supra* note 265 and accompanying text.

283. Daubert v. Merrell Dow Pharms., 509 U.S. 579 (1993).

284. *Id.* at 589.

285. Although, we can (potentially) know the methodology or principles used to build a proprietary computer system. This is still insufficient to ascertain that the system itself instantiates the model claimed for it. For physical devices like a gas chromatography apparatus, we can test the results by building our own using the known principles. Such reproducibility is not possible when the system, as in the case of mixture analysis, is proprietary.

that do not affect validation results but could affect outputs after mass deployment of the machine.

The only way to truly assess the model of the system is to gain access to the system code and then deconstruct it to its constituent parts to identify the model of decision-making employed to reach the final conclusion. If so, then we believe that system transparency is absolutely critical in assessing the admission of any Tier Two Mixed Model system and should be routinely ordered as a necessary prerequisite to proper reliability screening. As the National Institute of Science & Technology Report on DNA Mixture Analysis made clear, system design has not been forthcoming by the developers: "Currently, there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices, including the use of probabilistic genotyping software (PGS) systems."[286] We therefore agree with the approach of the New Jersey Superior Court in *State v. Pickett*, where the court ordered disclosure of the source code for TrueAllele to allow for independent assessment of the methodology the system uses to reach a conclusion.[287] Specifically, the court noted that authors of source code can be subject to "programmer's blindness" because "[j]ust as writers are often bad at proofreading their own text, programmers are bad at reading their own code . . . . It is often the case that peers are not truly independent reviewers because programmers often have similar training—and thus tend to make the same mistakes."[288] Only by gaining access to the system architecture can an independent evaluator determine the degree of reliability of the system, and thus admissibility in court. In this regard, the determination of the explainability of the system must be subsequent to opening the system for transparent review.

In summary, when faced with any machine-based expert evidence, the judge must be able to assess the underlying methodology of analysis because only then can the method be deemed reliable or not under *Daubert* or any other reliability screening test. For Tier One Known Models, the methodology of assessment will be transparent and, with guidance from the expert, can be traced from input data to the ultimate conclusion. The assessment by judges follows the model for any new scientific method, where the initial experts will be subjected to significant screening for reliability but, once established, the method will face fewer challenges. DNA evidence followed this pattern in the 1980s into the 1990s, but is

286. NAT'L INST. SCI. & TECH., DNA MIXTURE INTERPRETATION: A NIST SCIENTIFIC FOUNDATION REVIEW 6 (2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8351-draft.pdf [https://perma.cc/KHL2-8ENN].
287. State v. Pickett, 246 A.3d 279, 310 (N.J. Super. Ct. App. Div. 2021).
288. *Id.* at 314.

universally deemed reliable in modern litigation. Conversely, for Tier Three Machine Models, the methodology of assessment is not transparent, even to the operator, and as such, the methodology of analysis leading to a particular conclusion cannot be explained. If an expert witness in court could only testify to conclusions but not why they made them, the expert would quickly learn they would not be asked to testify often, as such *ipse dixit* opinions will be screened out after *Daubert* gatekeeping. The same is true for machine-based expert testimony. Finally, for Tier Two Mixed Models, the methodology of assessment is not transparent initially, but without it the judge is unable to assess the reliability of the decision-making process. If so, judges should routinely order disclosure of the underlying source code for Tier Two Mixed Models, allowing for independent assessment of the machine assisted decision process. This approach is not without precedent, as it has already been used in New Jersey to order source code disclosure for DNA mixture analysis of TrueAllele,[289] and as other Mixed Model machines become commercially available, a similar approach of transparency would be appropriate. After disclosure, the underlying methodology of analysis will result in expert testimony based upon it being deemed reliable or unreliable under *Daubert* or another gatekeeping test.

By ensuring transparency for all machine based expert testimony, the courts can appropriately screen next-level Mixed Model analysis for reliability ensuring that only reliable methodologies see the inside of a courtroom. This is the only way to ensure proper screening of machine-based evidence, and only way to ensure that, years from now, we do not face a similar reckoning as with hair, bite mark, and other discredited forensic methods.

B. Constitutional Considerations—Criminal Cases

Of course, reliability is not the sole consideration when offering evidence at trial. Criminal cases also contain constitutional constraints on the procedure of the trial and the evidence offered against the accused. To date, the main consideration for expert evidence analysis has been the confrontation clause right to cross examine witnesses,[290] but ultimately we find that right is unlikely to provide a meaningful review of the evidence beyond the reliability test discussed above.[291] However, due to the complexity of machine based expert testimony, we believe that considerations of due process will constrain litigants more by requiring

289. *Id.* at 310.
290. *See* discussion *infra* Section V.B.1.
291. *See* discussion *infra* Section V.B.2.

appointment of defense experts to review, analyze, and testify regarding complex machine based evidence.[292] Courts have generally been reluctant to appoint such experts, but considering the complexity, opacity, and specialization of Tier Two Mixed Model analysis, anything less than these appointments fails constitutional muster.

1. Confrontation: Unlikely to be a Significant Independent Constraint

Confrontation of witnesses against the accused is a core procedural right of the accused, and one that generally allows the defendant to ask questions of those witnesses in cross-examination to allow the jury to assess the witness's demeanor, memory, or overall credibility.[293] The Supreme Court extended the confrontation right to expert witnesses offering opinion evidence against the accused in a pair of cases, *Melendez-Diaz v. Massachusetts*[294] and *Bullcoming v. New Mexico*.[295] Initially, one may be tempted to suggest that the confrontation right would be a bulwark against improper machine based expert testimony, but upon further review we determine that it is largely duplicative of reliability analysis and, thus, of little use to screen machine testimony in court.

Imagine a Tier One Known Model analysis. In this type of assessment, an expert in, say, Gas Chromatography or Standard DNA analysis, can testify to the results of machine testing in the case, leading to an expert conclusion. In such a case, any testimony from the expert regarding the performance of the analysis or, if challenged, the reliability of the result, is likely to meet confrontation standards sufficient for Sixth Amendment purposes.[296] Thus, an expert testifying to the chemical test would testify to the mass spectrum of the sample and its pattern similarity to cocaine, leading to the conclusion that the substance is cocaine. In another case, the expert can show the results of standard capillary electrophoresis of the unknown DNA sample, leading to the conclusion that the unknown sample "matches" the known sample from the defendant. What both situations

292. *See* discussion *infra* Section V.B.2.

293. *See, e.g.*, United States v. Owens, 484 U.S. 554, 559 (1988) (finding confrontation has been satisfied when the defendant "has the opportunity to bring out such matters as the witness' bias, his lack of care and attentiveness, his poor eyesight, and even . . . the very fact that he has a bad memory"); California v. Green, 399 U.S. 149, 159–60 (1970) (rejecting a Confrontation Clause appeal since the witness was subject to cross for the jury, who could observe his demeanor and assess credibility of the witness).

294. 557 U.S. 305 (2009).

295. 564 U.S. 647 (2011).

296. *Id.* at 652 (noting that confrontation is satisfied when the defendant has the chance to cross the specific expert who did the analysis in the case).

would have in common is an explanation of the methodology of assessment leading to the conclusion, and, if challenged, a detailed explanation of the workings of the machine. Such cases rarely invoke confrontation challenges, as they facially meet the requirements of meaningful examination.

Conversely, consider a Tier Three Machine model consisting of a closed-box neural network. An expert relying on such a machine could take the stand and testify to their opinion, and if asked, explain that their opinion is based on the neural network or machine learning system they used. However, they could not explain how the machine went from inputs to the conclusion. If so, the testimony fails the most basic reliability screening and would be determined as an inadmissible *ipse dixit* expert opinion.[297] Even if the defendant wanted to cross the expert at trial, there would be no need as the testimony is, and should be, inadmissible as a matter of evidence law.

Lastly, consider a Tier Two Mixed Method analysis. Here, an expert testifies to the result of a mixture DNA analysis leading to a conclusion that the unknown sample contains defendant's allele pattern and therefore Defendant could be a contributor. Initially, the testimony may seem to be similar to Tier Three *ipse dixit* opinion, in that the analyst placed the sample in the machine and read the result to the court. However, there is a major difference between a Tier Two Mixed Method analysis and Tier Three *ipse dixit* opinion: if a Tier Two analysis is challenged, then the expert can explain the system design, show the system design, and explain the steps of assessment leading to the conclusion.[298] Thus, a Tier Two opinion can be shown to be reliable, if the initially unavailable transparency is ordered by the court as we suggest.[299] In this situation, the Confrontation Clause appears to provide no additional insight or benefit to the accused, as the testimony establishing the reliability of the machine analysis would more than suffice to meet the requirement of meaningful cross examination for Sixth Amendment purposes.

One word of caution is in order, regarding the underlying factual basis for the opinion. In the years after *Williams* and before *Smith*, some courts allowed experts to rely on the findings of a different expert in forming their own opinion because the underlying information would not be testimonial

---

297. Gen. Elec. Co. v. Joiner, 522 U.S. 136, 146 (1997) (stating ". . . nothing in either Daubert or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert. A court may conclude that there is simply too great an analytical gap between the data and the opinion proffered").
298. *See supra* Section IV.B.
299. *See* text accompanying notes 205–19.

and thus lacked confrontation protection after *Crawford*.[300] Some courts interpreted this rule to mean that any information that provided a proper basis of opinion under FRE 703 would not be testimonial, and therefore would lack protection from the Confrontation Clause.[301] *Smith* conclusively overruled that approach.

The *Smith* holding made clear that if an expert is relying on information as a basis of their opinion, and that the information only works as a basis of opinion if true, then the underlying statement is being offered for the truth and, thus, may be protected by confrontation rights.[302] In *Smith* itself, then, the statement of a second analyst about a chemical test could not be admitted when it relied on the actual testing of the first analyst, without confrontation.[303] And what would the Confrontation Clause require? The court stated that it must entail an evaluation of the truth of the underlying statement, because "[t]he jury cannot decide if the expert's opinion is credible without evaluating the truth of the factual assertions on which it is based."[304] Clearly, in the context of expert testimony based on scientific testing, what we need to know to assess the truth of the underlying assertion is whether the test was performed correctly and whether it reached an accurate result.[305] Confrontation allows that assessment as a constitutional matter.

On a practical level, though, the ramifications of *Smith* do not extend significantly beyond what is required already in the assessment of reliability. With a Tier Three model, for example, if the assessment of reliability cannot be done due to the opacity of the system model,[306] then there is no way to assess the truth of the assessment as a basis of testimony. *Smith*, then, would conclude that the confrontation right has been denied if a conclusion were offered. Yet in this example, the statement already failed reliability assessment, and if so, would not have been offered in the first place.[307] Imagine then a Tier One model, like in *Smith* itself, assessing a conclusion about gas chromatography. In *Smith*, now that confrontation applies, the court presumably would need to allow assessment of the truth of the underlying fact, which was the chemical test by the first analyst.[308]

---

300. Williams v. Illinois, 567 U.S. 50, 57–58 (2012), *abrogated in part by*, Smith v. Arizona, 602 U.S. 779 (2024).
301. United States v. Pablo, 696 F.3d 1280, 1287–88 (10th Cir. 2012), *overruled by*, Smith v. Arizona, 602 U.S. 779 (2024).
302. *Smith*, 602 U.S. at 795.
303. *Id.* at 796–99.
304. *Id.* at 796.
305. *Id.* at 798.
306. *See supra* Section V.A, text accompanying notes 254–59.
307. *See supra* Section V.A, text accompanying notes 254–59.
308. Smith v. Arizona, 602 U.S. 779, 790–91 (2024).

That truth, the ***validity*** of the underlying testing, is an assessment of the reliability of the machine analysis in question and, if so, would be satisfied by a recitation of the mechanism and its transparent, explainable process.[309] The reliability of the machine, then, establishes all requisite facts necessary for confrontation to be satisfied and, if reliability has been assessed as is required under *Daubert*, the confrontation right offers no additional protection.

Lastly, consider a Tier Two model. These systems differ from Tier One systems because they rely on complex statistical modeling that is not inherently transparent, and differ from Tier Three systems in that they are potentially transparent and explainable.[310] We concluded in Section V.A that the only way to assess these systems for reliability is to gain access to the system code, review its model of decision-making, and assess the reliability of the machine process in use.[311] Yet the implication of our approach to reliability is clear. If the assessment of reliability depends on a transparent assessment of the reliability of the mechanism of decision making in the machine, then any testimony that establishes that reliability would have to, on its face, meet the requirements of *Smith* for confrontation, namely an assessment of validity. If so, the Confrontation Clause should not offer any significant independent barrier to entry of testimony, so long as courts are screening machine systems for reliability properly.

For these reasons, we ultimately conclude that the Confrontation Clause fails to provide a significant independent barrier to admission of machine based expert testimony beyond the reliability screening necessary under *Daubert*. But that is not to say the constitution is silent on AI based expert evidence either.

2. Due Process: A Right to Expert Review for Tier Two Models

Even if confrontation is subsumed within reliability, we believe a more likely source of procedural protection for the accused in criminal cases can be found in the Fifth Amendment's Due Process Clause. In *Ake v. Oklahoma*, the Supreme Court discussed the obligation of the state to provide expert assistance in a criminal case, finding that an indigent defendant has the right to an expert as a matter of due process.[312] The court in *Ake* made clear that when the state charges an accused of a crime, it is necessary that they be provided a "fair opportunity to present [their]

---

309. *See supra* Section V.A, text accompanying notes 245–53.
310. *See generally* Section IV.B.
311. *See supra* text accompanying notes 260–69.
312. Ake v. Oklahoma, 470 U.S. 68 (1985).

defense."[313] Thus, under certain circumstances, an indigent defendant may be provided an expert at state expense, when the accused makes a showing that the expert is one of the "raw materials integral to the building of an effective defense."[314] In *Ake* itself, the accused could demonstrate that an expert psychiatrist was a necessary tool to assist in an insanity defense in a capital case, and failure to provide such assistance was a denial of due process.[315] Over time, courts have extended the holding of *Ake* to include non-capital cases [316] and experts beyond the psychiatric field.[317]

The exact contours of when an expert must be appointed have not yet been fully delineated, but the core concern is whether an accused is given a fair chance to rebut the charges in a case without an expert.[318] Thus, each case will require the court to assess the specific need for an expert to determine whether the evidence in question presents one of the "raw materials integral to building an effective defense."[319] As a matter of due process, courts have authorized experts in a wide variety of fields, including, among others: Odontology, Blood Spatter, DNA, Eyewitnesses, False Confessions, Pathology, Hypnosis, Pediatrics, Fingerprints, and Firearms.[320] Each case necessarily required the judge to assess the purpose of the expert within the context of the claims and defenses, finding trial would be unfair without the expert assistance.

---

313. *Id.* at 76.
314. *Id.* at 77.
315. *Id.* at 86–87,
316. Paul C. Giannelli, Ake v. Oklahoma: *The Right to Expert Assistance in a Post-*Daubert, *Post-DNA World*, 89 CORN. L. REV. 1305, 1365–68 (2004) (noting the expansion of *Ake* due process appointment of experts to areas beyond psychiatric experts, in a wide variety of fields). *See also* State v. Wang, 92 A.3d 220, 229 n.14 (Conn. 2014) (collecting Federal and State cases where experts were appointed in prosecutions for non-capital offenses).
317. Giannelli, *supra* note 316, at 1369–70 (finding that "[m]ost courts assume that *Ake* applies to noncapital cases") (citations omitted); Wang, 92 A.3d at 229 n.15 (collecting Federal and State cases on issue). *See also* Carlton Bailey, Ake v. Oklahoma *and an Indigent Defendant's 'Right' to an Expert Witness: A Promise Denied or Imagined?*, 10 WM. & MARY BILL RTS. J. 401, 437–40; 451–52 (2002) (discussing both federal and state cases where a court appointed a non-psychiatric expert for a defendant pursuant to due process).
318. Ake v. Oklahoma, 470 U.S. 68, 76 (1985).
319. *Id.* at 77.
320. *See, e.g.*, Little v. Armontrout, 835 F.2d 1240 (8th Cir. 1987) (hypnosis); Williams v. Martin, 618 F.2d 1021 (4th Cir. 1980) (pathology); Mentus v. Warden, No. 12-CV-447-JD, 2014 WL 2218676 (D.N.H. May 29, 2014) (firearms); People v. Warner, 22 N.W.3d 1 (Mich. 2024) (false confessions); Isham v. State, 161 So.3d 1076 (Miss. 2015) (pediatrics); Harrison v. State, 635 So.2d 894 (Miss. 1994) (odontology and pathology); James v. State, 613 N.E.2d 15 (Ind. 1993) (blood spatter); State v. Moore, 364 S.E.2d 648 (N.C. 1988) (fingerprints); State v. Bradley, 907 N.E.2d 1205 (Ohio Ct. App. 2009) (eyewitness identification); Dubose v. State, 662 So.2d 1156 (Ala. Crim. App. 1993) (DNA).

Of note to the topic of AI assisted expert testimony, courts have, as a matter of due process, authorized experts in computer forensics as well. For example, in *Lowe v. State*, the Mississippi Supreme Court overturned a conviction when it determined that the failure to provide a computer forensics expert to the defendant violated the *Ake* standard.[321] The case required not only that the state prove that the accused possessed the images on the computer but that he was the one who downloaded them, and to do so they relied on a prosecution expert in computer systems.[322] Under the circumstances, the court reasoned:

> Where, as here, the State relies on expert testimony alone to connect the defendant to the offense charged, an independent defense expert is part of the "raw materials integral to building an effective defense," and the trial judge deprives an indigent defendant of a fundamentally fair trial by refusing him funds to procure such an expert.[323]

The parallels of the *Lowe* court's reasoning to AI-assisted expert testimony are clear. In Part IV.A, we concluded that the disclosure of system architecture is necessary for assessment of the reliability of the machine-assisted evidence and should be routinely ordered with any evidence originating from a Tier Two Mixed Model system.[324] Only through this disclosure can an expert unwind the decision-making steps within the system, to understand how inputs to the system resulted in a final conclusion.[325] Yet this disclosure would not be of any assistance to any party unable to afford their own computer systems analyst. It is sheer folly to suggest that the systems are of such complexity that only disclosure for expert review would allow assessment of reliability, and then disallow expert appointments for that review. Thus, we conclude that in any case where expert system architecture is disclosed and the evidence from the machine model is central to the case (as is essential under *Ake*), that the court routinely order an expert to be appointed at state expense to review the machine architecture.[326] In the most common currently used Tier Two model, DNA Mixture analysis, the assessment of a program like

---

321. Lowe v. State, 127 So.3d 178, 184 (Miss. 2013).
322. *Id.* at 179–84.
323. *Id.* at 184.
324. See *supra* Section IV.A and text accompanying notes 245–48.
325. See *supra* Section IV.A and text accompanying notes 245–48.
326. Considering that the most common Type #2 Mixed Model system used in forensics today is DNA mixture analysis, it seems quite likely that the evidence will be central to the case in most if not nearly all cases when it is offered. *See supra* Section I.B and text accompanying notes 125–30.

STRMix or TrueAllele is certain to meet this test as central to the prosecution.[327]

It is clear that the appointment of due process experts in Tier Two Mixed Model cases will involve significant expense. Yet when considering expenses, one must also take into account the downside risk of not appointing the experts. In recent years, we have repeatedly seen convictions overturned based on faulty forensic evidence later deemed unreliable, in fields like bite marks, hair comparison, and other similar fields.[328] Had courts been instructed to examine these fields in detail at their conception, significant numbers of problematic convictions could have been avoided.[329] Now, with the development of AI-assisted evidence, the temptation is to immediately allow the evidence because the programmers believe it is reliable. Can we take them at their word? Considering the failures of other forensic fields, we cannot afford to take that risk and thus, smaller capital investments in individual cases are a prudent prophylactic against making similar mistakes again.[330] Economic benefits alone support this approach, independent of the inherent moral need for justice system legitimacy.

---

327. Kimberly Schweitzer & Narina Nunez, *What Evidence Matters to Jurors? The Prevalence and Importance of Different Homicide Trial Evidence to Mock Jurors*, 25 PSYCHIATRY, PSYCH. & L. 437, 443 (2018) (finding DNA evidence the most persuasive and important to mock jurors in deciding a hypothetical case); Laurie Meyers, *The Problem with DNA*, 38 MONITOR ON PSYCH. 52, 52 (2007) (quoting William Thompson, Ph.D., a prominent researcher in DNA since its beginnings in the 1980s, as stating "The problem with DNA is that it is perceived as providing a unique and infallible ID"). *See also* Giannelli, *supra* note 316, at 1391 (citing cases and commentators stating that, due to its complexity, DNA evidence should often lead to expert appointment) (citations omitted); John Devlin, *Genetics and Justice: An Indigent Defendant's Right to DNA Expert Assistance*, 1998 U. CHI. L. FORUM 395, 404 (1998) (stating that ". . . given the powerful impact of DNA evidence on a jury and the serious consequences involved, a defendant cannot enjoy a fair trial without his own DNA expert," in arguing for a due process right to a DNA expert in all DNA cases).

328. M. CHRIS FABRICANT, JUNK SCIENCE AND THE AMERICAN CRIMINAL JUSTICE SYSTEM 28 (2022). *See also* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 10, at 6–13 (explaining several fields that have been shown to be flawed, including microscopic hair comparison, bullet-lead evidence, footwear analysis, and bitemarks, before a comprehensive review of each field); NAT'L RSCH. COUNCIL, *supra* note 10, at 4 ("[I]n some cases, substantive information and testimony based on faulty forensic science analyses may have contributed to wrongful convictions of innocent people.").

329. NAT'L RSCH. COUNCIL, *supra* note 10, at 4 (noting that due to weaknesses in the scientific basis for many long-standing forensic science fields, "substantial improvement is necessary in the forensic science disciplines to enhance law enforcement's ability to identify those who have or have not committed a crime and to prevent the criminal justice system from erroneously convicting or exonerating the persons who come before it").

330. *Id.* at 20 (regarding the need for investment in research in forensics, the report sharply notes "[p]olitical and budgetary concerns should not deter bold, creative, and forward-looking action, because the country cannot afford to suffer the consequences of inaction").

Considering the complexities of machine evidence and the uncertainty of decision-making processes in Tier Two Mixed Models, we believe that computer forensic experts should be routinely appointed under *Ake* as a matter of due process, so long as the evidence in question is central to the prosecution's case, such as in current DNA mixture analysis using AI based systems.

VI. SUMMARY

At the cusp of a great technological change with the development of constantly improving AI systems, now is the time to stop to thoughtfully consider the use of these systems as the basis for expert testimony in criminal cases. AI-based expert testimony is subject to the constraints of reliability gatekeeping, and applying those rules largely depends on the system architecture of the AI system. Tier One models—"known models"—are based on established scientific principles, and are both transparent and reliable; if so, they can be shown to be reliable by testimony regarding their decision-making process. Standard DNA analysis and gas chromatography are examples of such methods and are universally accepted today in court. Tier Three models—Machine Models—are based on opaque processes of machine learning and will, over time, develop hidden layers of perceptions to analyze any problem. These systems are inherently non-transparent and, as such, cannot be tested nor evaluated for their decision-making method. If *Daubert*'s command that gatekeeping focus on "principles and methodology, not the conclusions they generate" has any meaning at all, then these systems are unreliable and inadmissible in court.[331]

Tier Two models—Mixed Models—show the current limits of the legal doctrines and can be admissible only under certain conditions. First, unlike Tier Three models, Mixed Models can be shown to be, although often are not inherently, transparent. If that is the case, courts must order disclosure of source code for these systems to allow for an independent assessment of the system decision-making processes. This disclosure, rather than the word of the programmers or output validation, is the sole way to assess if the machine works as designed. It is also the sole way to establish the reliability of the methodology used to reach a conclusion.

Additionally, constitutional limitations can also constrain the admission of AI-based expert evidence but not in the way commonly assumed. We believe that the confrontation right, one often examined as a potential constraint to machine evidence, is unlikely under current doctrine to serve as a meaningful constraint. What cases like *Melendez-Diaz* and

331. Daubert v. Merrell-Dow Pharms, 509 U.S. 579, 595 (1993).

Bullcoming demand is an opportunity to cross-examine an expert witness and ask meaningful questions to them about the system or test in question. However, since this testimony is undoubtedly required to establish the reliability of the system under FRE 702, it seems unlikely confrontation will provide additional limits to admission.

We do not conclude that the Constitution has nothing to say on the topic, however. Instead, we believe the source of constitutional doctrine to affect AI-based expert testimony admission is the due process clause. *Ake v. Oklahoma* commands that a defendant receive expert assistance at state expense when necessary to have an effective defense.[332] The complexity of AI system architecture almost goes without saying, and the disclosures we believe are mandated as a matter of reliability assessment would be of no use without a trained eye to evaluate the results. So, we believe that due process requires expert help to the accused, and is necessary in any situation with a Mixed Model serving as the basis of expert testimony.

The last decade or two have allowed the criminal justice system to reflect on past errors, reconsider long-standing beliefs, and challenge long-standing methods of proving guilt. Evidence once deemed routinely admissible has been challenged, successfully, on the grounds that it "lacked foundational validity."[333] As the criminal justice system sits on the verge of another revolutionary period of change, it would be foolhardy for courts to throw caution to the wind and deem AI-based evidence valid solely based on the word of the programmers. We believe that detailed assessment of system architecture, after appropriate disclosures, and assessment of that architecture by trained experts provided to the accused as a matter of due process is the way to ensure proper screening of evidence prior to admission, and should be the path forward.

---

332. Ake v. Oklahoma, 470 U.S. 68, 76 (1985).

333. PRESIDENT'S COUNCIL OF ADVISORS, *supra* note 10, at 13.