Artificial Intelligence and the Courts:

MATERIALS FOR JUDGES

Artificial Intelligence, Trustworthiness, and Litigation

SEPTEMBER 2022

AAAS
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Artificial Intelligence and the Courts: Materials for Judges

The American Association for the Advancement of Science (AAAS) is honored to have been entrusted, by the National Institute of Standards and Technology (NIST), with the task of developing educational materials on artificial intelligence (AI) for judges and courts.

AAAS therefore offers this compilation of educational materials for judges, covering a wide, yet appropriate, set of issues. (Please see the list below). AAAS' goal is to provide a set of user-friendly and accurate, yet readily comprehended, definitions, analyses and perspectives, on a variety of terms and topics with which the judiciary ought to become familiar.

The materials contained herein were developed by teams of scientific and legal experts who focused on a particular topic. The topics considered worthy of inclusion were selected based both on the mandate provided by NIST and guidance received by AAAS from an Advisory Committee composed of a large and diverse group of legal and AI experts. Drafts of the materials were subsequently submitted to Advisory Committee members, and outside expert "Reviewers," to obtain any suggestions for adjustments before each team of authors finalized their contribution (paper, podcast, annex, etc.).

It is not expected that courts will become experts regarding these sometimes complex or technical matters. Rather, this collection presents facts and overviews in a manner intended to make judges aware of key issues and to enable courts to find useful information contained herein, easily.

Finally, it is hoped that courts will appreciate certain innovative elements of this product, notably the inclusion of podcasts. These will provide courts with facts and analysis of important questions in a format that courts may find agreeable and, given the accompanying transcripts included, useful. AAAS thanks NIST for allowing a team of experts to undertake this forward-leaning approach to providing courts with needed information and insights as part of this project.

Materials in this series include:

1. ***Artificial Intelligence – Foundational Issues and Glossary***
2. ***Artificial Intelligence and the Justice System*** **(Podcast Series and Transcripts)**
   - Episode 1: *AI and Risk Scores* (49 minutes)
   - Episode 2: *AI in the Legal Field – Commercial and Unexpected Uses* (70 minutes)
   - Episode 3: *AI, Decision-Making, and the Role of Judges* (58 minutes)
3. ***Artificial Intelligence, Trustworthiness, and Litigation***
4. ***Artificial Intelligence, Legal Research, and Judicial Analytics***
5. ***Artificial Intelligence and Bias – An Evaluation***

Acknowledgements

The American Association for the Advancement of Science (AAAS) gratefully acknowledges the authorship of Cynthia Cwik, Senior Legal Advisor, Former Stanford Fellow, and former Partner at Jones Day and Latham & Watkins; Paul W. Grimm, Judge, US District Court for the District of Maryland; Maura Grossman, Research Professor, David R. Cheriton School of Computer Science at the University of Waterloo; and Toby Walsh, Scientia Professor of Artificial Intelligence at the University of New South Wales.

We also extend our gratitude to the members of the Advisory Committee for their input and guidance throughout the design and production of this product: Rediet Abebe, Assistant Professor, Computer Sciences, University of California, Berkeley; Theodore F. Claypoole, Partner, Womble, Bond, Dickenson, LLP; John Cooke, Director, Federal Judicial Center; Cynthia Cwik, Senior Legal Advisor, Former Stanford Fellow, and Former Partner, Jones Day and Latham & Watkins; Herbert B. Dixon, Jr., Judge (ret.), Superior Court of the District of Columbia, and Senior Legal Advisor to the Center for Legal and Court Technology, William & Mary School of Law; Finale Doshi-Velez, Associate Professor, Computer Science, School of Engineering and Applied Sciences, Harvard University; Paul W. Grimm, Judge, US District Court for the District of Maryland; James Harris, Technology Consultant, National Center for State Courts; Joseph Sawyer, Director, Faculty Development and Online Learning, National Judicial College; Elham Tabassi, Chief of Staff, Information Technology Laboratory, National Institute of Standards and Technology; and Zachary Zarnow, Principal Court Management Consultant, National Center for State Courts.

AAAS appreciates the contributions of Solomon Assefa, Vice President, IBM Research; Nicole Clark, Chief Executive Officer, Trellis Research; Rebecca Crootof, Associate Professor of Law, University of Richmond School of Law; Patrick Huston, Brigadier General and Assistant Judge Advocate General for Military Law and Operations, US Department of Defense; Maia Levy-Daniel, Attorney and Public Policy Consultant; Matthew Stepka, Managing Partner, Machina Ventures, and Lecturer, Haas School of Business, University of California, Berkeley; Peter Stone, Professor, Department of Computer Sciences, and the Founder and Director, Learning Agents Research Group, Artificial Intelligence Laboratory, University of Texas, Austin; Rebecca Wexler, Assistant Professor of Law and the Faculty Co-Director, Berkeley Center for Law & Technology, University of California, Berkeley; and others for their review and insightful comments on previous drafts of the *Artificial Intelligence and the Courts: Materials for Judges* series.

This paper was edited by Joel Ericsen, Program Associate, and Alain Norman, Senior Program Associate, Science and the Law, AAAS. Special thanks to Theresa Harris, Program Director, Scientific Responsibility, Human Rights and Law Program, and Nathaniel Weisenberg, Program Associate, AAAS, and to Barbara “Bebe” Holloway, University of Virginia, for their contributions.

AAAS Creative and Brand Manager, Paula Fry, created the cover design for this publication.

The production of this publication and other materials in the *Artificial Intelligence and the Courts: Materials for Judges* series was funded by the United States National Institute of Standards and Technology (NIST) through Award 60NANB21D031.

Disclaimer: The opinions, findings and conclusions or recommendations expressed in this publication do not necessarily reflect the views of the AAAS Board of Directors, its council and membership, or NIST.

Contact: AAAS welcomes comments and questions regarding its work. Please send information, suggestions and any comments to the AAAS Scientific Responsibility, Human Rights and Law Program at srhrl@aaas.org.

Cite as: Cwik, C., Grimm, P., Grossman, M. and Walsh, T. "Artificial Intelligence, Trustworthiness, and Litigation." *Artificial Intelligence and the Courts: Materials for Judges*. American Association for the Advancement of Science (2022). Available at https://doi.org/10.1126/aaas.adf0786.

Abstract

Although few court decisions have squarely addressed the admissibility of artificial intelligence (AI) evidence in proceedings governed by the Federal Rules of Evidence, or their state-law equivalents, this paper focuses on key considerations for the use of AI evidence in court cases. The paper defines the concept of "trustworthiness" as being the sum total of a number of interrelated requirements found within the rules of evidence that govern court cases. This section also includes:

- *Annex A*: "Practice Pointers for Lawyers and Judges," given the complexities and rapid evolution of AI, this Annex offers a handy set of practical questions courts might employ, the better to assess the validity, reliability and/or admissibility of proffered AI-related evidence.
- *Annex B*: "Hypothetical on the Admissibility of Facial Recognition Testimony in a Criminal Matter," provides a fact-pattern and legal framework for analyzing a scenario of the sort that a court might plausibly encounter.
- *Annex C*: "Hypothetical on Measuring a Machine Learning (ML) System's Accuracy and Reliability—Problem Gambling," provides a fact-pattern an Australian court has encountered, as well as sample questions for any court needing to assess ML-related issues.

Table of Contents

Artificial Intelligence, Trustworthiness, and Litigation

Authors:

Cynthia Cwik
Senior Legal Advisor
Former Stanford Fellow
Former Partner, Jones Day and Latham & Watkins

Maura Grossman
Research Professor, David R. Cheriton School of Computer Science, University of Waterloo

Paul W. Grimm
Judge, US District Court for the District of Maryland

Toby Walsh
Scientia Professor of Artificial Intelligence, University of New South Wales

1. Introduction

As **artificial intelligence**[1] (AI) applications[2] become more ubiquitous in different aspects of our lives, it seems unavoidable that the evidence needed to resolve civil litigation and criminal trials will include outputs that are generated by this rapidly evolving technology. Thus, lawyers seeking to introduce or object to AI evidence, and judges who must rule on its admissibility, must have a basic knowledge of what AI is and how it works, and how to evaluate its trustworthiness. This is because, with AI—**machine learning** (ML) in particular—questions about the data on which it was trained (including its representativeness of the population on which the AI will be used), the inner workings of the **algorithm** (including its features and weights) and how the output was derived can all be difficult to explain to judges and juries lacking a background in computer or data science. This can create challenges when evaluating the *trustworthiness* of AI evidence, which, in the context of court cases, means its relevance, **validity**, **reliability** and **authenticity**. Because this section focuses on the use of AI evidence in court cases, we will define the concept of "trustworthiness" as being the sum total of a number of interrelated requirements found within the rules of evidence that govern court cases. For the purposes of this section, AI evidence is sufficiently trustworthy to be introduced into evidence when it meets the requirements of the rules of evidence.[3]

---

[1] Bolded red terms appear in the Glossary

[2] Paul W. Grimm, Maura R. Grossman, and Gordon V. Cormack, *Artificial Intelligence as Evidence*, 19 Nw. J. Tech. & Intell. Prop. 9 (2021). Available at: https://scholarlycommons.law.northwestern.edu/njtip/vol19/iss1/2/.

By the term "AI," we mean to refer to computer systems and applications that are capable of performing functions normally associated with human intelligence, such as abstracting, reasoning, problem solving, learning, etc. See *AI as Evidence* at 14-17. Such systems may use one or more algorithms, including, but not limited to, rules-based systems, machine learning, natural language processing, deep learning, and machine vision. While at times in this section we may appear to be referring solely to systems that use machine learning—systems that are "trained" to recognize patterns in data and to derive models that can explain the data or make predictions about other data—this is by way of example, only, and by no means intended as a limitation.

[3] *See AI as Evidence* at 84-97.

There are few, if any, published court opinions that consider issues involving AI admissibility in any depth. Recently, however, governments and other organizations have been working on proposed AI governance frameworks, with the goal of mitigating the risks of AI, and these efforts can provide useful guidance. For example, the U.S. Department of Commerce's National Institute of Standards and Technology ("NIST") is developing an AI Risk Management Framework, to provide guidance regarding the trustworthiness of **AI systems**.[4] Specifically, the framework is intended to help to incorporate trustworthiness considerations into the design, development, use and evaluation of AI systems. These trustworthiness characteristics include "accuracy, **explainability** and **interpretability**, reliability, **privacy**, robustness, safety, security (resilience) and mitigation of unintended and/or harmful **bias**, as well as of harmful uses."[5] Once completed, the NIST framework will likely influence how companies and other organizations approach AI-related risks, and may provide useful context for judges and practitioners concerning AI design and uses when evidence generated by AI-powered software is introduced or objected to in court cases.

For judges who must decide whether to admit AI evidence, it is important to determine the validity of an AI application (i.e., how accurately the AI measures, classifies, or predicts what it is designed to), as well as its reliability (i.e., the consistency with which AI produces accurate results when applied in the same or substantially similar circumstances).[6] Factors that can affect the validity and reliability of AI evidence, include bias of various types, lack of **transparency** and explainability and the sufficiency of the objective testing of the AI application before it is released for public use.[7] Closely related to the problem of inadequate testing and evaluation is the problem of **function creep**, which refers to the gradual widening of the use of a technology or system beyond the use for which it was originally intended, often, but not always, without its proper validation for the new use.[8]

With AI evidence, the significance of validity and reliability, and the factors that impact it, can be different than with other types of evidence. For example, although explainability is often considered to be important when evaluating the validity and reliability of evidence, different considerations may be necessary when evaluating AI evidence, which may be a "**black box**,"[9] or may involve an immense number of data points. *See, e.g.,* K. Miller, *Should AI Models be Explainable? That Depends*, Stanford HAI News (March 16, 2021) (noting that **AI models** that

---

[4] NIST, *AI Risk Management Framework Concept Paper*, December 13, 2021, https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf.

[5] *See id.*

[6] See *AI as Evidence* at 32 n.92, 49-51, 79-83, 98-99.

[7] *See id.* at 13-14, 42-47, 48-50, 60-65.

[8] *See id.* at 51-53. *See also function creep*, dictionary.com, https://www.dictionary.com/browse/function-creep (last visited December 16, 2021).

[9] "In science, computing, and engineering, a black box is a device, system, or object which can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings." Will Kenton, *Black Box Model*, Investopedia, https://www.investopedia.com/terms/b/blackbox.asp (last visited Apr. 24, 2022).

yield accurate predictions that help clinicians better treat their patients can be useful even without a detailed explanation of how or why the models work).

The following subsection will discuss issues that frequently arise during the pretrial phase of litigation (i.e., the discovery phase), where the parties exchange information about the facts that are relevant to resolving the issues raised by the pleadings or charges that have been filed with the court in the case. It will provide an overview of the evidentiary principles that govern whether AI evidence should be admitted in court cases. The focus of this discussion is on providing a step-by-step analysis of the most important issues, and the factors that affect decisions on whether or not to admit AI evidence. The accompanying Annex A includes a summary of practical suggestions intended to assist lawyers and judges as they are called upon to introduce, object to, or decide on whether to admit AI evidence. In Annex B, we provide a hypothetical example involving the admissibility of **facial recognition** technology in a criminal matter, with a discussion of the relevant rules and factors to consider. Finally, Annex C, based on an actual case in Australia, provides sample questions courts anywhere might wish to leverage in cases involving machine learning.

2. Admissibility Issues

The Federal Rules of Evidence[10] are amended infrequently, and the process of amendment is slow. In contrast, technology, and especially AI technology, changes at near-breakneck speed, and often is incorporated into routine use by individuals, organizations, corporations and governments long before it is the subject of evidentiary scrutiny in a particular case. However, the Federal Rules of Evidence are resilient and are designed to be used in a manner that is flexible. Rule 102 provides: "These rules should be construed so as to administer every proceeding fairly, eliminating unjustifiable expense and delay, and *promote the development of evidence law*, to the end of ascertaining the truth and securing a just determination" (emphasis added).[11] Thus, we believe, the existing Federal Rules of Evidence are adequate for the task of evaluating AI evidence, provided they are applied flexibly.

Relevance and authenticity are the two areas that create most of the evidentiary challenges for admitting AI evidence, and they are the main focus of this subsection.[12] Other evidence

[10] Every state in the United States has adopted its own rules of evidence, some of which are identical or nearly identical to the Federal Rules of Evidence, and some of which differ in significant respects. Nonetheless, the evidentiary concepts that govern admissibility of AI evidence are fundamental, and found in all compilations of the rules of evidence. Further, the Federal Rules of Evidence are frequently cited as persuasive authority even in states that have evidence codes that differ from the Federal Rules. For that reason, this section will focus on the Federal Rules of Evidence because of their national scope and their influence on state codifications of the rules of evidence. *See AI as Evidence* at 84 & n.333.

[11] Fed. R. Evid. 102.

[12] *See AI as Evidence* at 85.

doctrines, such as the hearsay rule,[13] and the original writing rule,[14] can be encountered, but these rules present less of a concern than authenticity. The focus of the hearsay rule is intentionally assertive statements made by human declarants,[15] and AI applications, by their very nature, involve machine-generated output.[16] While the evidence may, and often does, take the form of an express or implied factual assertion (e.g., "this is the photo of the person depicted in the surveillance video"; "this is the sector of the city that is likely to have the greatest potential for criminal activity on a particular day and time;" "this job applicant is likely to be the most qualified for the vacancy being filled"), and may be offered for its substantive truth, the source is not a *human* declarant, therefore it is not properly regarded as hearsay.[17] Rather, the key issue is *authenticity*—how accurately does the AI system that generated the evidence produce the result that its proponent claims it does. Similarly, the original writing rule imposes a requirement that proof of the content of writings, recordings and photographs must be made by introducing an original or duplicate original,[18] but those terms are defined interchangeably, and broadly, so they are seldom difficult to comply with, unless a witness is called who merely describes what he or she observed as the output of the AI system, instead of introducing a copy.[19] This seldom occurs for the simple reason that having a human describe the contents of the output of an AI system that produces a written, recorded, or photographic result robs it of most of the weight that the evidence would have if the jury were shown the output itself (once properly authenticated).[20]

2.1. Relevance

Federal Rule of Evidence 401 defines *relevance*. It states: "Evidence is relevant if: (a) it has any tendency to make a fact more or less probable than it would be without the evidence; and (b) the fact is of consequence in determining the action." This is a relatively low bar for admitting evidence, because even evidence that has only a slight tendency to prove or disprove facts that

[13] *See* Fed. R. Evid. 801-807.

[14] *See* Fed. R. Evid. 1001-1008.

[15] *See* Fed. R. Evid. 801(a)-(c).

[16] "Because human design, input, and operation are integral to a machine's credibility, some courts and scholars have reasoned that a human is the true 'declarant' of any machine conveyance. But while a designer or operator might be partially epistemically or morally responsible for a machine's statements, the human is not the sole source of the claim… . The machine is influenced by others, but is still a source whose credibility is at issue." Andrea Roth, *Machine Testimony*, 127 Yale L.J. 1972, 1977-78 (2017). *See also AI as Evidence* at 85-86 & n.340.

[17] *See, e.g., U.S.* v. *Wallace*, 753 F.3d 671, 675 (7th Cir. 2014) (rejecting confrontation clause challenge to the admissibility of a video recording showing an exchange of drugs between two people because there was no human declarant to be cross examined and there was no showing that the conduct involved was intended by the participants to be an assertion, therefore, there was no hearsay "statement," as contemplated by Fed. R. Evid. 801(a), and no "declarant," as contemplated by Fed. R. Evid. 801(b); *U.S.* v. *Lizarraga-Tirado*, 789 F. 3d 1107, 1109-10 (9th Cir. 2015) (rejecting hearsay challenge to a satellite image and accompanying GPS coordinates).

[18] *See* Fed R. Evid. 1001 (defining duplicates and duplicate originals), 1002 (setting forth the substantive rule), and 1004-1007 (setting forth exceptions to the rule).

[19] *See AI as Evidence* at 86.

[20] *See id.*

are important to resolving a civil or criminal case can meet this standard.[21] Examined in isolation, it could be argued that AI evidence that has not adequately been examined to determine its validity and reliability still has some tendency to prove a disputed issue. Rule 401 does not require perfection, only a tendency to prove or disprove.[22]

Rule 401 must be considered along with Rules 402 and 403. Rule 402 states: "Relevant evidence is admissible unless any of the following provides otherwise: the United States Constitution; a federal statute; these rules [of evidence]; or other rules prescribed by the Supreme Court. Irrelevant evidence is not admissible."[23] In essence, Rule 402 creates a presumption that relevant evidence is admissible, even if it is only minimally probative, unless other rules of evidence or sources of law require its exclusion.[24] While the first part of Rule 402 is flexible, the second part is immutable: irrelevant evidence is never admissible.[25]

Rule 403 provides: "The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time or needlessly presenting cumulative evidence."[26] As it relates to the admissibility of AI evidence, Rule 403 has three important features. First, it establishes a "balancing test" for determining whether relevant evidence may be considered by the judge or jury. This scale "tilts" towards admissibility of relevant evidence.[27] Such evidence is inadmissible only if its probative value (*i.e.*, its ability to prove or disprove important facts presented in a case) is *substantially outweighed* by the adverse consequences listed in the rule. It is not enough that relevant evidence will be prejudicial to the party against which it is introduced—after all, all evidence offered by a plaintiff or the government against a defendant is intended to be prejudicial in the sense that it is offered to show that the defendant is liable or guilty. It is excludable only if its prejudice is *unfair* to that party.[28] Similarly, Rule 403 will tolerate a degree of confusion on the part of the judge or jury that must evaluate the evidence, even if it might mislead them, provided that these adverse consequences do not substantially outweigh the tendency of the evidence to prove important facts in the case.[29] Even though the balancing in Rule 403 favors admissibility,

---

[21] *See id. at* 86-87. *See* also Michael M. Martin, Stephen A. Salzburg, and Daniel Capra, 1 *Federal Rules of Evidence Manual* § 402.02[1], at 401-6-7 (12th ed. 2019) ("To be relevant it is enough that the evidence has a *tendency* to make a consequential fact even the least bit more probable or less probable than it would be without the evidence). (emphasis in original)).

[22] *See AI as Evidence* at 87.

[23] Fed. R. Evid. 402.

[24] *See AI as Evidence* at 87.

[25] *See id*.

[26] Fed. R. Evid. 403.

[27] *See, e.g.*, *United States v. Terzado-Madruga*, 897 F. 2d 1099, 1117 (11th Cir. 1990) (holding that the balancing test of Fed. R. Evid. 403 "should be struck in favor of admissibility.").

[28] *See United States* v. *Guzman-Montanez*, 756 F.3d 1, 7 (1st Cir. 2014) ("[T]he law shields a defendant against unfair prejudice not against all prejudice. '[A]ll evidence is meant to be prejudicial; it is only unfair prejudice which must be avoided.'"). *See also AI as Evidence* at 87-88.

[29] *See id.* at 88.

the fact that the rule clearly establishes that judges must consider unfairness, must be aware that confusion may result, and must be careful to discern whether the jury may be misled, is extremely important, especially when applied to the admissibility of AI evidence.[30] After all, the court cannot evaluate technical evidence for prejudice, confusion, or assess whether it misleads without some understanding of how it works.[31] Similarly, judges cannot assess whether a jury will be misled or confused by AI evidence unless they have an appreciation for whether the AI application meets acceptable standards of validity and reliability, which may differ depending on what the evidence is being offered to prove, and the adverse consequences flowing from allowing a jury composed of lay persons to consider that evidence in reaching its verdict.[32]

Second, Rule 403 makes it clear that the trial judge acts as a gatekeeper, charged with the responsibility of reviewing the evidence, in the first instance, to determine whether the jury may hear it.[33] This obligation flows from another rule of evidence, such as Rule 104(a), which states: "The court must decide any preliminary question about whether a witness is qualified, a privilege exists, or evidence is admissible. In so deciding, the court is not bound by evidence rules, except those on privilege."[34] Implicit in this delegation of responsibility to the court is the notion that the judge must have the tools to make this preliminary determination.[35] The hallmark feature of the American justice system is that it is an adversary process, and so it is the responsibility of the parties, not the judge, to develop and present the factual evidence that will be offered to the jury for its consideration.[36] Accordingly, lawyers who intend to offer (or challenge) AI evidence must do the work necessary to explain to the judge how the AI system works (including, for example, how it was programmed or trained, how it operates, and how it produced its output), why the evidence will enlighten not confuse and how it will promote a just outcome, not one that is unfair.[37]

Because of the technical complexity of AI evidence, the trial judge must raise with the parties, well in advance of the trial, the question of whether they intend to offer AI or similarly technical evidence at trial, and as part of the pretrial scheduling process, impose reasonable deadlines for disclosing an intention to introduce such evidence, and for challenging its admissibility, sufficiently far in advance of trial to allow the judge to have a hearing (which will likely require the testimony of expert witnesses).[38] Determinations about whether AI evidence meets adequate thresholds of validity and reliability sufficient for it to be considered by the jury do

[30] *See id.*

[31] *See id.*

[32] *See id.*

[33] *See id.*

[34] See *id.*

[35] *See id.*

[36] *See id.*

[37] *See id.* at 89.

[38] *See id.*

not lend themselves to last minute, on-the-fly assessments, and should not be attempted or allowed in the middle of a trial itself.[39]

Finally, it should be obvious that a judge cannot make the determinations required by Rules 401 through 403 unless the party offering the AI evidence is prepared to disclose underlying information concerning, for example, the **training data** (if any) and the development and operation of the AI system sufficient to allow the opposing party (and the judge) to evaluate it, and the party against whom the AI evidence will be offered to decide whether and how to challenge it.[40] If a party intends to rely on output that is the product of an AI application in a civil or criminal trial, they should not be permitted to withhold from the party against whom that evidence will be offered the information necessary to determine the validity (i.e., the degree of accuracy with which the AI system measures what it purports to measure), and the reliability (i.e., the consistency with which the AI system correctly measures what it purports to measure under similar circumstances), of the AI evidence.[41] If they are prohibited from doing so by claims of proprietary information or trade secrets raised by the company that developed the AI application, the trial judge should consider giving the proponent of the AI evidence a choice: either disclose the underlying evidence (subject to an appropriate protective order), or otherwise demonstrate its validity and reliability.[42] If the proponent is unwilling or unable to do so, then serious consideration should be given as to whether they should be precluded from introducing the AI evidence at trial.[43]

In sum, invalid or unreliable AI systems produce results that have insufficient tendency to prove or disprove disputed facts in a trial and/or that are unduly prejudicial. Neither the trial judge nor the party against whom AI evidence is offered should be required to accept at face value the unproven claims of the proponent of the evidence that it is valid and reliable.[44]

2.2. Authentication of AI Evidence

Federal Rule of Evidence 901(a) sets forth, in plain terms, what is meant by the requirement that AI evidence must be *authenticated* in order to be considered by the jury. It states: "To satisfy the requirement of authenticating… an item of evidence, the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is."[45] Rule 901(b) then lists 10 non-exclusive ways in which a party can accomplish this task.[46] The examples that most readily lend themselves to authenticating AI evidence are: Rule 901(b)(1) (testimony of a witness with knowledge that an item is what it is claimed to be); and Rule

[39] *See Id.*

[40] *See id*.

[41] *See id.*

[42] *See id.*

[43] *See id.*

[44] *See id.* at 90.

[45] Fed. R. Evid. 901(a). *See also AI as Evidence* at 90.

[46] *See* Fed. R. Evid. 901(b)(1)-(10). *See also AI as Evidence* at 90.

901(b)(9) (evidence describing a process or system and showing that it produces an accurate result).[47]

When authenticating AI evidence using Rule 901(b)(1), the testimony of the witness called to perform this task must comply with other rules of evidence.[48] For example, Rule 602 requires that the authenticating witness have personal knowledge of how the AI technology functions.[49] It states: "A witness may testify to a matter only if evidence is introduced sufficient to support a finding that the witness has personal knowledge of the matter. Evidence to prove personal knowledge may consist of the witness's own testimony. This rule does not apply to a witness's expert testimony under Rule 703." [50]

There are some important features of Rule 602 that tend to be overlooked by some lawyers and judges.[51] There is an understandable tendency to call the fewest possible number of witnesses to authenticate evidence.[52] When a single person possesses all the knowledge needed to do so, then that is all that is required.[53] However, AI applications seldom are the product of a single person possessing personal knowledge of all the facts that are needed to demonstrate that the data used as input, the technology itself, and its output are what its proponent claims them to be. Data scientists may be required to describe the data used to train an AI system using machine learning.[54] Developers may be required to explain the features and weights that were chosen for the machine-learning algorithm.[55] Technicians knowledgeable about how to operate the AI system may be needed to explain what they did when they used the tool, and the results that they obtained.[56] These technicians, however, may not be able to explain how the data was collected or cleansed, how the machine-learning algorithm that underlies the system was trained, or how the system was tested to show that it produces valid

[47] *See id.* at 91.

[48] *See id.*

[49] *See* Charles A. Wright and Victor J. Gold, 31 *Federal Practice and Procedure: Evidence* §7103, at 24-25 (2000), which states that "[f]or purposes of analyzing the scope of Rule 901, the most important additional relationship is the one between that provision and Rule 602… . Both Rules 602 and 901 identify elemental qualities that make evidence worthy of consideration. Since the provisions perform similar functions, it is important to know when evidence is subject to the personal knowledge requirement of Rule 602 and when it is subject to the authentication or identification requirement of Rule 901. Rule 602 applies only to testimonial evidence… . Rule 901 does not apply to testimonial evidence, it applies to all other evidence. The distinction can be misleading, however, because it might be taken to suggest that Rule 602 and 901 never apply to the same evidence. In fact, these provisions are simultaneously applied where testimony is the means by which some respect of non-testimonial evidence is relayed to the jury." *See also AI as Evidence* at 91.

[50] Fed. R. Evid. 602.

[51] *See AI as Evidence* at 91.

[52] *See id.*

[53] *See id.*

[54] *See id.*

[55] *See id.*

[56] *See id.*

and reliable results.[57] Still other witnesses may be needed to interpret the output of the AI system.[58]

Rule 702 provides that: "A witness who is qualified as an expert by knowledge, skill, experience training or education may testify in the form of an opinion or otherwise if (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case."[59]

Importantly, Rule 703 states that: "An expert may base an opinion on facts or data in the case that the expert has been made aware of or personally observed. If experts in the particular field would reasonably rely on those kinds of facts or data in forming an opinion on the subject, they need not be admissible for the opinion to be admitted."[60] If the requirements of Rules 702 and 703 were met, then, a party that wanted to authenticate an AI system that was developed by a team of individuals with scientific, technical, or specialized knowledge beyond the personal knowledge of any one person could do so with a single qualified expert.[61] However, the requirements of Rules 702 and 703 are quite demanding when applied as intended by the Federal Rules of Evidence.[62]

In sum, lawyers must bear in mind, and judges must be vigilant to require, that the witness or witnesses called to authenticate AI evidence either have personal knowledge of the authenticating facts or qualify as an expert that is permitted to incorporate into their testimony information from sources beyond their own personal knowledge, provided it is sufficiently reliable.[63]

The second authenticating rule most suited to AI evidence is Rule 901(b)(9).[64] It permits authentication by "[e]vidence describing a process or system and showing that it produces an accurate result."[65] To do so, the party that wishes to introduce the AI evidence would face the same challenges just described in the discussion of Rule 901(b)(1)—calling a single witness or

[57] *See id.*

[58] *See id.*

[59] Fed. R. Evid. 702.

[60] Fed. R. Evid. 703. *See also AI as Evidence* at 93.

[61] *See id.*

[62] *See id.*

[63] *See*, *e.g.*, Fed. R. Evid. 703. *See also United States* v. *Frazier*, 387 F. 3d 1244, 1260 (11th Cir. 2004) (discussing the importance of a trial judge diligently fulfilling his or her gatekeeping function under Fed. R. Evid. 104(a) to ensure the "reliability and relevancy of expert testimony" because an expert's opinion "can be both powerful and quite misleading because of the difficulty in evaluating it."). *See also AI as Evidence* at 93.

[64] *See id*.

[65] Fed. R. Evid. 901(b)(9).

witnesses themselves possessing personal knowledge of all the authenticating facts, or qualifying as an expert under Rules 702 and 703.[66]

An important feature of authentication needs careful consideration in connection with admitting AI evidence.[67] Normally, a party has fulfilled its obligation to authenticate non-testimonial evidence by producing facts that are sufficient for a reasonable factfinder to conclude that the evidence more likely than not is what the proponent claims it is.[68] In other words, by a mere preponderance. This is a relatively low threshold—51%, or slightly better than a coin toss.[69] However, not all AI evidence is created equal.[70] Some AI systems have been independently tested and shown to be valid and reliable.[71] Others have not, when, for example, efforts to obtain information sufficient to test their validity and reliability have been blocked by claims of proprietary information or trade secret.[72] Moreover, some of the tasks for which AI applications have been put to use can have serious adverse consequences if they do not perform as promised—such as arresting and criminally charging a person based on flawed facial recognition technology, or sentencing a defendant to an extended term of imprisonment based on a machine-learning system that has been trained using biased or incomplete data that inaccurately or differentially predicts the likelihood that the individual will reoffend.[73]

The greater the risk of unacceptable adverse consequences, the greater the need to show that the AI system is unlikely to produce those consequences.[74] Judges, tasked with making the initial determination of admissibility of AI evidence under Rule 104(a), should be skeptical of

---

[66] There are two additional rules of evidence that may be used to authenticate AI evidence that are closely related to Rules 901(b)(1) and 901(b)(9). They are Fed. R. Evid. 902(13), which allows authentication of "[a] record generated by an electronic process or system that produces an accurate result, as shown by a certification of a qualified person"; and Fed. R. Evid. 902(14), which allows authentication of "[d]ata copied from an electronic device, storage medium, or file, if authenticated by a process of digital identification, as shown by a certification of a qualified person." Rules 902(13) and (14) would allow the proponent of AI evidence to authenticate it by substituting the certificate of a qualified witness for their live testimony. However, the qualifications of the certifying witness and the details of the certification that the evidence produces an accurate and reliable result must be the same as would be required by the in-court testimony of a similarly qualified witness. *See* Charles A. Wright and Victor J. Gold, *supra* n.49 §7147, at 43, stating that "[n]ewly adopted Rule 902(13)] allows the authenticity foundation that satisfies Rule 901(b)(9) [process or system producing accurate results] to be established by a certification rather than the testimony of a live witness. If the certification provides information that would be insufficient to authenticate the record if the certifying person testified, then authenticity is not established under Rule 902(13)." The same applies for the certification in Rule 902(14), certified data copied from an electronic device, storage medium, or file. *See AI as Evidence* at 93.

[67] *See id.* at 94.

[68] *See Lorraine v. Markel Am. Ins. Co.*, *supra* n.325 at 542; *United States v. Safavian*, 435 F. Supp. 2d. 28, 38 (D.D.C. 2006); *United States v. Holmquist*, 36 F. 3d 154, 168 (1st Cir. 1994) ("the standard for authentication, and hence admissibility, is one of reasonable likelihood."). *See also AI as Evidence* at 94.

[69] *See id.*

[70] *See id.*

[71] *See id.*

[72] *See id.*

[73] *See Id.*

[74] *See id.*

admitting AI evidence that has not been shown to be accurate by much more than an evidentiary coin toss.[75] They should insist that the proponent of the evidence establish the validity and reliability of the AI to a degree that is commensurate with the risk of the adverse consequences likely to occur if the technology does not perform as claimed.[76] If the proponent of the evidence fails to do so, then the trial judge should evaluate under Rule 403 whether the probative value of AI authenticated by a mere preponderance is substantially outweighed by the danger of unfair prejudice to the adverse party or would confuse or mislead the jury to an unacceptable degree,[77] taking into consideration the nature of the adverse consequences that could occur if the AI system is insufficiently valid or reliable.[78]

2.3. *Daubert* Factors and the Admissibility of Expert Evidence

Federal Rule of Evidence 702 requires that introduction of evidence dealing with scientific, technical, or specialized knowledge that is beyond the understanding of lay jurors be based on a sufficient facts or data and reliable methodology that has been applied reliably to the facts of the particular case.[79] These factors were added to the Federal Rules of Evidence in 2000 to bolster them in light of the U.S. Supreme Court's decisions in *Daubert* v. *Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), and *Kumho Tire Co.* v. *Carmichael*, 119 S. Ct. 1167 (1999).[80] Therefore, while Rule 702 was not intended to codify the *Daubert* decision, the factors discussed in that decision relating to determining the reliability of scientific or technical evidence are quite informative when determining whether Rule 702's reliability requirement has been met.[81] As described in the Advisory Committee Note to the amendment of Rule 702 that went into effect in 2000, the "*Daubert* Factors" are: "(1) whether the expert's technique or theory can be or has been tested…; (2) whether the technique or theory has been subject to peer review and publication; (3) the known or potential rate of error of the technique or theory when applied; (4) the existence and maintenance of standards and controls; and (5) whether the technique or theory has been generally accepted in the scientific [or technical] community."[82] Most state courts have also adopted some version of the *Daubert* factors when considering the admissibility of scientific evidence.

---

[75] *See id.*

[76] *See id*.

[77] *See* Fed. R. Evid. 403. *See also AI as Evidence* at 94-95.

[78] *See id.* at 95.

[79] *See* Fed. R. Evid. 702 (b)-(d). *See also generally In re Paoli R.R. Yard PCB Litig.*, 35 F. 3d 717, 742 (3d Cir. 1994) (discussing the importance of the reliability factor in the *Daubert* analysis, and the obligation of the trial judge to "take into account" all of the factors listed in *Daubert* that are relevant to determining the reliability of the scientific or technical evidence that is being offered into evidence). *See also AI as Evidence* at 95.

[80] *See* Advisory Committee Note, Fed. R. Evid. 702 (2000).

[81] *See AI as Evidence* at 95 & n.369. It should be noted that when the term "reliability" is used in the Federal Rules of Evidence and related case law, it encompasses both the scientific notions of validity (*i.e.*, accuracy) and reliability (*i.e.*, consistency under substantially similar circumstances).

[82] *See* Advisory Committee Note, Fed. R. Evid. 702 (2000).

Using the *Daubert* factors, in order to authenticate AI evidence, its proponent must show that it produces valid (meaning accurate) results.[83] It also must perform reliably, meaning that it consistently produces accurate results when applied in substantially similar circumstances.[84] When the validity and reliability of AI evidence has been verified through independent testing and evaluation of the AI system that produced it, the methodology used to develop the evidence has been published and subject to review by others in the same field of science or technology, when the error rate associated with the AI system is not unacceptably high, when standard methods and protocols for operation of the AI system have been followed, and when the methodology used is generally accepted within the field of similar scientists or technologists, then it has been authenticated.[85] It does what its proponents say it does.[86] And introducing evidence from such a system or application produces none of the adverse consequences against which Rule 403 is designed to guard.[87]

In contrast, when the validity and reliability of a system or process that produces AI evidence has not properly been tested, when its underlying methodology has been treated as a trade secret by its developer preventing it from being independently verified by others, when applying the method produces unacceptably high error rates, when corners were cut and standard procedures were not followed when the system was developed or employed, or when the methodology is not accepted as valid and reliable by others in the same field, then it is hard to say that it does what its proponent claims it does, which ought to render it inauthentic and inadmissible.[88] The bottom line is that if a lawyer intends to rely on AI evidence to prove their case, they should consider these five *Daubert* factors and marshal the facts to show compliance with as many of them as they can.[89] Courts should insist that the party offering evidence produced by an AI system to prove its case adequately show that it does what its proponent claims it does, to a degree of certainty commensurate with the risk of an unacceptably bad outcome if it turns out that the technology is unreliable.[90] Failing that, the AI evidence should be excluded for insufficiency of authentication (under Rule 901(a)), failure to show the use of reliable methodology that was applied to the facts of the case (under Rule 702), and/or excessive danger of unfair prejudice, or of confusing or misleading the jury (under Rule 403).[91]

[83] *See AI as Evidence* at 96.

[84] *See id.*

[85] *See id.*

[86] *See id.*

[87] *See id.*

[88] *See id.*

[89] *See id.*

[90] *See id.*

[91] *See id.* at 96-97.

3. Conclusion

Although the adoption of AI within an increasingly large sector of society is a relatively recent development, it is undoubtedly here to stay.[92] AI is in a state of such rapid advancement that the law of evidence governing the circumstances under which AI systems and their output should be admitted into evidence in civil and criminal trials is not well developed.[93] Although some commentators have written about potential problems and concerns that impact whether AI evidence should be admitted, there are few court decisions that have squarely addressed the admissibility of AI evidence in proceedings governed by the Federal Rules of Evidence or their state-law equivalents.[94] This will change over time, as it is inevitable that AI systems and their inputs and outputs will be at the center of disputes that will increasingly find their way into court.[95] When this happens, lawyers and judges must be prepared to address the evidentiary issues that influence whether the AI evidence should be admitted.[96] Since AI systems are complex and highly technical, most lawyers and judges will be ill equipped for this task unless they have at least a rudimentary understanding of what AI is, how it operates, methods of scientific and statistical evaluation that impact decisions about its validity and reliability, and hence, its admissibility.[97] Because there are at present no rules in the Federal Rules of Evidence that directly address AI evidence, lawyers and judges must rely on the rules that do exist to provide an analytical framework to assist them when they confront these issues.[98]

[92] *See id.* at 105.

[93] *See id.*

[94] *See id.*

[95] *See id.*

[96] *See id.*

[97] *See id.*

[98] *See id.*

Annex A: Practice Pointers for Lawyers and Judges

If lawyers and judges accept the fact that there are myriad types and uses of AI, and that there are many potential challenges raised by AI—for example, potentially risk of bias, lack of robust testing and validation, function creep, lack of transparency and explainability, and lack of resilience—all of which can all affect the validity and reliability of AI evidence—and they recognize the need to authenticate AI evidence properly before it is admitted into evidence (following the rules that govern how to do so), then the question arises: How should lawyers faced with introducing or challenging AI evidence, and judges who must rule on its admissibility, go about doing so? Below, we offer some practical suggestions with the hope that they will make this task less daunting in practice.[99]

A.1. What was the AI Designed to Address?

The essence of much AI technology, particularly that which relies on ML, comes down to:

1. the data used to train the system;
2. the algorithm(s) which comprise the system (including, but not limited to, their **features**, weights and operation); and
3. the models, predictions, or outputs that result from running the system.[100]

Algorithms are simply a set of rules or procedures for solving a problem or accomplishing an end.[101] So, the starting point for determining the admissibility of AI technology is to understand the problem that the AI was designed to solve.[102] Knowing this is essential to assessing:

1. the appropriateness of the data used to train the system, and whether it is representative of the data on which the system will be used;
2. the validity of the system (*i.e.*, its accuracy in performing the intended function);
3. its reliability (i.e., the consistency with which it produces the same or substantially similar results when applied under substantially similar circumstances); and
4. whether it is being used for purposes for which it was not designed (i.e., whether there has been substantial function creep).[103]

The proponent of the evidence should start with the AI's design objective in order to begin to amass the evidence necessary to secure its admissibility.[104] Opposing parties need to know this

[99] *See AI as Evidence* at 97.

[100] *See id.*

[101] *See id.*

[102] *See id.*

[103] *See id.*

[104] *See id.*

information to be able to intelligently assess whether its admissibility should be challenged.[105] And judges need to know this to be able to rule on the admissibility of the evidence derived from the AI system.[106] Relevance is not an abstract concept. Evidence is relevant only to the extent that it has the ability to prove or disprove facts that are consequential to the resolution of a case. The problem that the AI was designed to address—and the output it produces—must "fit" with what is at issue in the litigation.[107] Without knowing what the AI was designed and programmed to do, none of these fundamental questions can begin to be answered.[108]

A.2. How was the AI Developed and by Whom?

One of the issues that affects the validity and reliability of AI evidence is whether its design was influenced by improper bias, whether intended or not.[109] Was the data used to train the system skewed or complete?[110] Is it representative of the target population on which the system will be used?[111] If the AI system was trained with historical data that reflects discrimination, how was this addressed? Were variables incorporated that are proxies for impermissible characteristics (*e.g.*, zip code or arrest records, which may correlate with and therefore incorporate race)?[112] What assumptions, norms, rules, or values were used to develop the system? Were the people who did the programming themselves sufficiently qualified, experienced and/or diverse to ensure that there was not inadvertent bias that could impact the output of the system?[113] Did the programmers given due consideration to the population that will be affected by the performance of the system?[114] These questions cannot be answered without knowledge of certain factors, including information about the data that was used as input for purposes of training, how the AI system was developed and by whom, including the design choices that were made, how the system was operated and how the output was produced and interpreted.[115] Judges should be particularly careful not to allow a party planning to introduce AI evidence to hide behind claims of proprietary information or trade secrets without careful consideration of the consequence to the party against whom the AI evidence will be offered.[116] Will allowing trade-secret claims to shield disclosure of how the AI system was developed, trained and functions prevent the party against whom it will be introduced from having a fair opportunity to learn how the AI works (and where it may have defects) so

[105] *See id.*

[106] *See id.*

[107] *See id.*

[108] *See id.*

[109] *See id.* at 98.

[110] *See id.*

[111] *See id.*

[112] *See id.*

[113] *See id.*

[114] *See id.*

[115] *See id.*

[116] *See id.*

that they can prepare a defense?[117] If so, how are they to frame evidentiary challenges to its use?[118] Adverse parties who are refused access to the information they need to assess AI's validity and reliability on the basis of claims of trade secrets should challenge these designations and seek a ruling from the court that either grants them access to the information they reasonably need (subject to proper protective measures) or prohibits the introduction of the AI evidence at trial.[119] Judges must ask themselves how they can fulfill their gatekeeping role in ruling on the admissibility of the AI evidence if presented with little more than a black-box AI program and a conclusory claim that it is accurate and consistently functions as it was designed to.[120]

A.3. Were the Validity and Reliability of the AI Sufficiently Tested?

Validity and reliability are key concepts in assessing whether AI evidence should be admitted as evidence.[121] The proponent of AI evidence should be required to demonstrate that the AI system that produced the evidence being offered has been tested (preferably independently) to confirm that it is both valid for the purpose for which it is being offered, and reliable.[122] If it was not tested, why not, and on what basis is the proponent claiming that it operates as intended, and consistently so?[123] And why should the court even consider allowing the introduction of the output of an untested AI system?[124] Who designed and carried out the testing?[125] Was it the same people who developed the system in the first place?[126] If so, was the methodology used to test the system standard or otherwise reasonable, adhering to procedures accepted as appropriate by the relevant scientific or technical community familiar with the subject matter at the heart of the AI system?[127] Under what conditions did the testing occur and how to they compare to the circumstances under which the system is now being used?[128] Was the system tested for both validity and reliability?[129] Has the validity and reliability been confirmed by others who are independent of the developers?[130] Are the results of the testing still available so that they may be reviewed by the adverse party and the court?[131]

---

[117] *See id.*

[118] *See id.*

[119] *See id.*

[120] *See id.*

[121] *See id.*

[122] *See id.* at 98-99.

[123] *See id.* at 99.

[124] *See id.*

[125] *See id.*

[126] *See id.*

[127] *See id.*

[128] *See id.*

[129] *See id.*

[130] *See id.*

[131] *See id.*

The answers to these questions should inform the court's decision as to whether the evidence should be admitted at all.[132] Allowing the introduction of AI evidence derived from a system that has not been shown to be valid and reliable for the purpose for which the evidence is being introduced substantially increases the risk that its probative value (if any) is substantially outweighed by the danger of unfairly confusing or misleading the factfinder.[133] This is particularly the case if the AI evidence is the primary evidence being offered to prove an essential element of the proponent's case.[134]

A.4. Is the Manner in Which the AI Operates "Explainable" So that It Can be Understood by Counsel, the Court and the Jury?

An important factor in evaluating the admissibility of AI evidence is whether the functioning of the AI system that produced the evidence can be explained to the trier of fact, who may be unfamiliar with the technology and methodology involved, so they can understand, at least at a general level, how the system operates, how it achieves its results, and thus, evaluate the amount of weight they are willing to give to the evidence derived from it.[135] NIST offers useful guidance in this regard in its publication titled *Four Principles of Explainable Artificial Intelligence*. [136] The NIST authors describe four principles of explainable AI:

- **Explanation**: Systems deliver accompanying evidence or reason(s) for all outputs.
- **Meaningful**: Systems provide explanations that are understandable to individual users.
- **Explanation Accuracy**: The explanation correctly reflects the system's process for generating the output; and
- **Knowledge Limits**: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output.[137]

Although written from the perspective of scientists interested in the development and/or evaluation of valid and reliable AI methods, the discussion emphasizes the same themes that underlie the purpose of the rules of evidence: that when technical information is offered during a trial, the proponent of that evidence must demonstrate that it is sufficiently trustworthy for the trier of fact to credit it in making its decision.[138] If the proponent of the evidence cannot even explain how the AI system operates in a way that can be understood by the trier of fact (including assuring them that it is only being used under the conditions for which it was designed, describing the system's error rate, and showing that there is acceptable confidence in

[132] *See id.*

[133] *See id.*

[134] *See id.*

[135] *See id.*

[136] *See* P. Jonathon Phillips et al., *Four Principles of Explainable Artificial Intelligence*, NIST Pub. NISTIR 8312 (Sept. 2021).

[137] *Id.* at ii. *See also AI as Evidence* at 99-100.

[138] *See id.* at 100.

its accuracy), that can affect whether the evidence produced from the system should be admitted by the court.[139]

A.5. What is the Risk of Harm if AI Evidence that is Not Shown to be Trustworthy is Admitted?

The Federal Rules of Evidence do not require that all risk of error be eliminated before scientific and technical evidence may be admitted.[140] Evidence is relevant if it has any tendency, however slight, to prove or disprove facts that are important to deciding a case.[141] And authenticity is established if the proponent demonstrates that the evidence more likely than not is what it purports to be.[142] The argument could be made that even AI evidence shown to be valid and reliable for a particular purpose, but which is being offered to prove something for which its validity and reliability have not been established, may have some tendency to prove what it is being offered to prove.[143]

The expert witness rules[144] are helpful for evaluating the admissibility of AI evidence because they supply demanding standards:

1. whether there is a sufficient factual basis to support the evidence;
2. whether the methods and principles used to generate the evidence were reliable; and
3. whether they were reliably applied to the facts of the particular case.[145]

The *Daubert* factors further focus the inquiry on the following:

1. whether the methodology was tested;
2. whether there is a known error rate;
3. whether the methods used are generally accepted as reliable within the relevant scientific or technical community that is familiar with the methodology;
4. whether the methodology has been subject to peer review by others knowledgeable in the field; and
5. whether standard procedures or protocols are applicable to the methodology, and if they were complied with.[146]

---

[139] *See id.*

[140] *See id.* at 101.

[141] *See* Fed. R. Evid. 402. *See also AI as Evidence* at 101.

[142] *See id.*

[143] *See id.*

[144] *See* Fed. R. Evid. 702; 703.

[145] *See* Fed. R. Evid. 702. *See also AI as Evidence* at 101.

[146] *See Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 593-94 (1993). *See also AI as Evidence* at 101.

But even this enhanced level of analysis does not require perfection.[147] The ultimate question that must be decided in each case is *whether the evidence is sufficiently valid and reliable for the purpose for which it is being offered*. The answer to this question will depend on what is at stake if the fact finder credits AI evidence that is invalid and unreliable.[148]

A.6. Timing Issues

Determining whether AI evidence should be admitted at trial is complicated, requires a great deal of information and is not the type of issue that is well suited to being resolved in the middle of a trial, or on the fly.[149] Preparation is critical, both by the proponent and opponent of the AI evidence.[150] The judge needs time to hear the competing evidence, to carefully review the supporting materials and to decide.[151] But since there is no rule of evidence that specifically addresses AI evidence, nor do the Federal Rules of Civil or Criminal Procedure directly require the disclosure of AI evidence, there is a risk that it may not be disclosed soon enough for disputes about its admissibility to be determined before trial.[152]

It is true that a party that intends to call a witness who would meet the definition of an expert witness under Fed. R. Evid. 702, in order to lay the foundation for AI evidence, would have to disclose the witnesses' opinions and the basis therefore, which should give its adversary and the court some advanced notice that AI evidence is going to be introduced.[153] But expert disclosures often are more generally about the subjects of the expert's intended testimony than the rules actually require, such that the intent to introduce AI evidence may not be clearly flagged far enough ahead of trial.[154] That means that the parties should communicate well ahead of trial to determine whether AI evidence is going to be offered at trial, and reach agreement (or bring the matter to the attention of the court) about when such AI evidence will be disclosed, the extent to which the party against whom the AI evidence will be proffered will have access to the information needed to assess and challenge its validity and reliability, and whether the proponent of the AI evidence will assert proprietary information or trade-secret protection to deny the production of such information to the opposing party.[155]

The trial judge should also inquire during the pretrial stage of the case whether AI evidence will be introduced, set a deadline for its production, as well as for challenges to its admissibility, rule on any trade-secret claims and schedule a hearing well before trial to ensure that the court itself is adequately informed and has sufficient time to make a principled decision as far in

[147] *See id.*

[148] *See id.*

[149] *See id.*

[150] *See id.* at 104.

[151] *See id.*

[152] *See id.*

[153] *See* Fed. R. Civ. P. 26(b)(4); Fed. R. Cr. P. 16(a)(1)(G). *See also AI as Evidence* at 104.

[154] *See id.*

[155] *See id.* at 105.

advance of trial as possible.[156] Finally, a trial judge faced with ruling on the admissibility of AI evidence need not rely solely on the arguments of the attorneys for the parties and their experts but can appoint a court expert as permitted by Fed. R. Evid. 706,[157] if the circumstances so warrant.[158]

[156] *See id.*

[157] *See* Fed. R. Evid. 706. *See also AI as Evidence* at 105.

[158] *See id.*

Annex B: Hypothetical on the Admissibility of Facial Recognition Testimony in a Criminal Matter

B.1. Factual Background

Defendant Jamal Warner has been charged with armed robbery, assault and brandishing a firearm in the Meridian County Circuit Court, State of South Sunland. Since his arrest in October, 2021, he has been held in pretrial detention. He is represented by an attorney in the South Sunland Public Defender's Office. An Assistant District Attorney for Meridian County is the prosecutor.

The indictment alleges that on August 21, 2021, at 8:45 PM, Warner, wearing a hoodie with the hood pulled up and sunglasses, entered the Deluxe Jewelry Store shortly before closing time. He produced a handgun, and ordered the only employee present, Bob Parker, the store manager, to put all of the cash in the register and in the store safe into a gym bag, along with all the high-end jewelry. Warner brandished the firearm as he demanded the cash and jewelry, threatened to shoot Parker, and when Parker dropped some jewelry on the counter, Warner hit him on the side of his head with the firearm. Warner then grabbed the gym bag and fled the store. The scene was captured on the store's surveillance video, which is grainy and slightly out of focus. While it is possible to see the robber's actions, his facial features are partially obscured by his hoodie and the sunglasses, and the angle at which the camera is pointing makes it difficult to determine Warner's height. It can be determined, however, that he is a dark-skinned African American male, with a close-cropped beard, who appears to be of thin build. Parker, the store manager, is a 57-year-old white male.

Meridian County police officers responded to the scene minutes after Warner fled the store, alerted by the alarm that went off when activated by Parker as Warner was fleeing. They obtained a copy of the surveillance video, which was given to Investigator Mary Adams, a digital forensic examiner, who viewed it. Adams, who also is white, selected three still frames from the video that showed three-quarters of Warner's partially turned head more clearly than any other frames of the video. She then loaded these three images into a forensic facial recognition software program that the Meridian Police have licensed from its manufacturer, Accu-Match. Then, using the Accu-Match program, she accessed the South Sunland State Central Criminal Records Database, she scanned the booking photographs of all Black males in that database. All of these photos are face-on photos, taken under good lighting conditions. The Accu-Match software uses an AI algorithm to compare exemplar digital images to a survey set of digital images contained in the database being surveyed. Adams followed the steps she learned when she was trained how to use the Accu-Match software to run the three images taken from the surveillance video against the booking photographs in the Central Criminal Records database. This search resulted in 52 "matches" that were produced in the following categories: highly probable match (15 photos), probable match (17 photos) and possible match (20 photos).

Adams selected five photos from the "highly probable match" photos that Adams thought most closely resembled the images in the jewelry store video. All five were African American males with beards. She arranged these five photos in a photo-array, showed them to Parker, who studied them carefully before saying "It's hard to tell, because the robber was wearing dark glasses and a hoodie, but I'm pretty sure it was photo number three." Photo number three was a booking photo of Warner taken in May 2015, when he was arrested for drunk and disorderly conduct. On the basis of that identification, Adams obtained an arrest warrant, and Warner was arrested, charged with robbery, assault and brandishing a firearm, and detained while awaiting trial.

Warner's Public Defender has filed a motion to suppress the pretrial identification of Warner. An evidentiary hearing on this motion has been scheduled by Circuit Court Judge Gail Langley. Under the South Sunland Rules of Criminal Procedure, the rules of evidence govern pretrial suppression motions in criminal cases. The South Sunland Rules of Evidence are identical to the Federal Rules of Evidence. Prior to the motion's hearing Warner's attorney requested the issuance of a subpoena to the Accu-Match Company to compel them to produce the Accu-Match software and its source code, so that a digital forensic examiner hired by counsel for Warner can examine and test it, to determine how it functions and its accuracy. The prosecutor objected to the issuance of the subpoena, and counsel for Accu-Match filed a motion to quash the subpoena. They both argued that the source code of the Accu-Match was proprietary, confidential trade-secret information that should not be produced in discovery. However, the prosecutor proffered to Judge Langley that it would authenticate the Accu-Match software with an appropriate witness that would establish its accuracy. Judge Langley granted the motion to quash, and declined to issue the subpoena.

Thirty days before the evidentiary hearing the prosecutor filed with the court and served on the Defendant a Certification signed under penalty of perjury by Investigator Adams, attached to which were copies of the three images of the robber taken from the jewelry store surveillance video, and the five Central Criminal Records images that were selected from among the "highly probable match" set produced by the Accu-Match AI. The Certification was made pursuant to South Sunland Evidence Rule 902(13), which permits the authentication of records generated by an electronic system or process shown to produce accurate results. In the Declaration, Adams stated that she had been a police officer in the Meridian County Police Department for 17 years, five years as a patrol officer, seven years as a detective in the violent crimes division and five years as a digital forensic examiner. With respect to her qualifications as a digital forensic examiner, Adams' declaration stated that she had attended a nine-month forensic examiner training course at the South Sunland Law Enforcement Academy (where she learned how to extract digital information from digital devices, desktop computers, laptops, tablets and smart phones), followed by two years as an assistant forensic examiner, during which time she worked along with a senior forensic examiner on actual cases, and received further on-the-job-training in forensic examination. Two years earlier she was selected to attend a three-month training course sponsored by Accu-Match, where she was trained in how to operate its AI

software to perform facial recognition examinations comparing exemplar digital facial images to a comparison set of digital images. At the conclusion of that training, she was certified as a Accu-Match examiner by the company. She outlined the step-by-step procedures required when using the Accu-Match software, and confirmed that she followed each step as trained to do. In addition, she stated that she had been using this software for more than 18 months in dozens of criminal investigations, and that in each case, the software produced highly probable matches that resulted in arrests and in many of those cases criminal charges had been issued. Finally, she stated that in each case in which she used the Accu-Match software, her selection results were peer-reviewed by another certified digital forensic examiner in her office who also was a certified Accu-Match examiner. Finally, she stated that she had testified in three trials as to her use of this software in making a facial recognition match, had been qualified as an expert in each instance, and the evidence of her selections was admitted into evidence at trial, where the defendant was convicted.

B.2. Framework for Legal Issues Regarding the Admissibility of the Accu-Match Facial Recognition Software

B.2.(a). Relevance Rules of Evidence

- **Federal Rule of Evidence 401:** "Evidence is relevant if: (a) it has any tendency to make a fact more or less probable than it would be without the evidence; and (b) the fact is of consequence in determining the action." This is a relatively low bar to admitting evidence.
- **Federal Rule of Evidence 402:** "Relevant evidence is admissible unless any of the following provides otherwise: the United States Constitution; a federal statute; these rules [of evidence]; or other rules prescribed by the Supreme Court. Irrelevant evidence is not admissible." In essence, Rule 402 creates a presumption that relevant evidence is admissible, even if it is only minimally probative, unless other rules of evidence or sources of law require its exclusion.
- **Federal Rule of Evidence 403:** "The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time or needlessly presenting cumulative evidence." As it relates to the admissibility of AI evidence, Rule 403 establishes a "balancing test" for determining whether relevant evidence may be considered by the judge or jury. It is inadmissible only if its probative value (*i.e.*, its ability to prove or disprove important facts presented in a case) is substantially outweighed by the adverse consequences listed in the rule. Similarly, Rule 403 will tolerate a degree of confusion on the part of the judge or jury that must evaluate the evidence, even if it might mislead them, provided that these adverse consequences do not substantially outweigh the tendency of the evidence to prove important facts in the case. Even though the balancing in Rule 403 favors admissibility, the fact that the rule clearly establishes that judges must consider unfairness, be aware

that confusion may result, and be careful to discern whether the jury may be misled, is extremely important, especially when applied to the admissibility of AI evidence. Similarly, judges cannot assess whether a jury will be misled or confused by AI evidence unless they have an appreciation for whether the AI application meets acceptable standards of validity and reliability, which may differ depending on what the evidence is being offered to prove, and the adverse consequences flowing from allowing a jury composed of lay persons to consider that evidence in reaching its verdict.

- **Federal Rule of Evidence 104(a):** "The court must decide any preliminary question about whether a witness is qualified, a privilege exists, or evidence is admissible. In so deciding, the court is not bound by evidence rules, except those on privilege." Lawyers who intend to offer (or challenge) AI evidence must do the work necessary to explain to the judge how the AI system works (i.e., produced its output), why the evidence will enlighten not confuse, and promote a just outcome, not one that is unfair.

B.2.(b). Authenticity Rules of Evidence

- **Federal Rule of Evidence 901(a):** "To satisfy the requirement of authenticating … an item of evidence, the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is." Rule 901(b) lists 10 non-exclusive ways in which a party can accomplish this task. The examples that most readily lend themselves to authenticating AI evidence are: Rule 901(b)(1) (testimony of a witness with knowledge that an item is what it is claimed to be); and Rule 901(b)(9) (evidence describing a process or system and showing that it produces an accurate result).
- **Federal Rule of Evidence 901(b)(9):** It permits authentication by "[e]vidence describing a process or system and showing that it produces an accurate result." To do so, the party that wishes to introduce the AI evidence can call a single person or persons themselves possessing personal knowledge of all the authenticating facts or qualifying as an expert under Rules 702 and 703.
- **Federal Rule of Evidence 902(13):** This rule allows for self-authentication of "[a] record generated by an electronic process or system that produces an accurate result, as shown by a certificate of a qualified person that complies with the certification requirements of Rule 902(11) or (12). The proponent must also meet the notice requirement of Rule 902(11).

B.2.(c). Witnesses

- **Federal Rule of Evidence 602:** "A witness may testify to a matter only if evidence is introduced sufficient to support a finding that the witness has personal knowledge of the matter. Evidence to prove personal knowledge may consist of the witness's own testimony. This rule does not apply to a witness's expert testimony under Rule 703.

B.2.(d). Rule 702 and the 'Daubert Factors' Regarding the Admissibility of Expert Testimony

- **Federal Rule of Evidence 702**: "A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:
    1. the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
    2. the testimony is based on sufficient facts or data;
    3. the testimony is the product of reliable principles and methods; and
    4. the expert has reliably applied the principles and methods to the facts of the case."
- **'*Daubert* Factors'**: The factors discussed in the U.S. Supreme Court's decisions in *Daubert* v. *Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), and *Kumho Tire Co.* v. *Carmichael*, 119 S. Ct. 1167 (1999) relating to determining the reliability of scientific or technical evidence are informative when determining whether Rule 702's reliability requirement has been met. As described in the Advisory Committee Note to the amendment of Rule 702 that went into effect in 2000, the "*Daubert* Factors" are:
    1. "whether the expert's technique or theory can be or has been tested…;
    2. whether the technique or theory has been subject to peer review and publication;
    3. the known or potential rate of error of the technique or theory when applied;
    4. the existence and maintenance of standards and controls; and
    5. whether the technique or theory has been generally accepted in the scientific [or technical] community."[159]

B.3. Specific Factual Considerations with Respect to the Admissibility of the Accu-Match Facial Recognition Software

Factors relating to the reliability and quality of probe photos from the Deluxe Jewelry Store video:

- Resolution;
- Lighting;
- Distance of the suspect from the camera;
- Orientation of probe photo (*i.e.*, facial angle);
- Occlusion of face with mask, glasses, facial hair, hoodie or hat etc.;
- Facial expression of suspect;

[159] *See* Advisory Committee Note, Fed. R. Evid. 702 (2000).

- Demographics for suspect (*e.g.*, race, gender, age);
- Any editing of probe photos;
- Number of probe photos that were not used with software and reason for excluding those photos.

Factors relating to photo database:

- Origin of photos, including how they were selected and by whom;
- Age of photos;
- Resolution;
- Lighting;
- Any editing of photos;
- Number of photos in database of individuals with similar characteristics to suspect in terms of:
    - Distance of the suspect from the camera;
    - Orientation of probe photo (i.e., facial angle);
    - Occlusion of face with mask, glasses, facial hair etc.;
    - Facial expression of suspect;
    - Demographics for suspect (i.e., race, gender, age).

Factors relating to Accu-Match software:

- Known error rate or bias (i.e., training data was not sufficiently representative of exemplars similar in demographics to Defendant Warner or algorithm has higher error rate with certain demographics);
- Validation studies, including with regard to individuals with a similar demographic background to Defendant Warner and whether those studies were conducted independently or by Accu-Match itself;
- Proficiency tests;
- Software updates;
- Peer-reviewed literature relating to this or similar software;
- Industry standards or controls;
- General acceptance of this specific type of technology and the particular algorithm used in the scientific community;
- Ability to test software, including using source code.

Factors relating to Investigator Adams' testimony:

- Knowledge, skills, training and education regarding facial recognition software generally, and Accu-Match software specifically (in other words, does she have the specialized knowledge or skill to testify to the validity and reliability of the software itself, or is her knowledge limited to her training and experience regarding how to use the software, in

which case she would not be qualified to provide the certification under Rule 902(13) establishing that the product of using the software was the result of a system or process that produced an accurate result);

- Specific procedures used in this matter to make the match with Defendant Warner;
- Demographic considerations, including similarity with Defendant Warner and examiner's potential biases;
- Specific experience of the digital forensic examiner peer-reviewer with Accu-Match software, and demographics regarding peer-reviewer, including potential biases;
- Consideration of the demographics of Bob Parker, the store manager, and potential biases.

B.4. Final Thoughts

1. In deciding the admissibility of the evidence of the Accu-Match identification, the presiding judge must first determine whether it has been properly authenticated by Investigator Adams. Although she provided a certificate to authenticate the fact that the results produced by Accu-Match were the result of a system or process that produces accurate results (*i.e.*, the standard articulated by Rules 901(b)(9) and 902(13), does Adams have the training, knowledge and experience to testify either form persona knowledge or expertise as to how the software was developed, trained and tested (all of which require expertise), or is she merely relaying conclusory statements told to her when she was trained on how to use the software? In other words, is she the correct person to authenticate this evidence?
2. The trial judge must resolve the issue of whether the defense attorney should be given access to source code or other information about how the Accu-Match system operates, to be able to independently test it to have a basis to challenge its accuracy. While this information may be a trade secret or confidential proprietary information of Accu-Match, that does not render it immune from discovery, and an outright prohibition of discovery to confirm the software's accuracy may raise due process issues. A better approach is to allow reasonable discovery by the defense, subject to a protective order.
3. Finally, after considering all the evidence in favor of and against admitting the Accu-Match photo match, is the judge satisfied that the software is sufficiently valid and reliable (*i.e.*, the result of a system or process that produces accurate results) to outweigh the danger of unfair prejudice that would result from an identification that is based on insufficiently accurate evidence? The judge would not only consider the identification match generated by Accu-Match, but also the strength of Parker's identification, the actual security video, the three images selected by Investigator Adams to use with Accu-Match, the selections made by Accu-Match, the selection of the five photos from the 52 Accu-Match "matches," and whether the Defendant's attorney has had a fair opportunity to receive discovery sufficient to challenge the accuracy of the Accu-Match software.

Annex C: Hypothetical on Measuring a Machine Learning System's Accuracy and Reliability—Problem Gambling

C.1. Forword

For judges who must decide whether to admit evidence, it is important to determine the accuracy and reliability of an AI system under inspection. The following example illustrates some of the challenges in doing this. It is adapted from the author's recent experience as an expert witness in a case in Australia, and has been modified to protect identities.

C.2. Fact Pattern

As a responsible corporation, the Emerald Casino contracted Daedalus Research to build a Machine Learning (ML) system to identify problem gamblers on their slot machines. The system was to take various inputs such as bet size, bet timing and bet frequency, as well as personal information extracted from video cameras such as gender and estimated age. The ML system was then required to classify a person using a slot machine into one of two classes: problem or non-problem gambler.

Daedalus Research built a system to perform this classification and delivered it to the Emerald Casino. However, the matter ended up in the courts when the Emerald Casino refused to pay for the system, disputing the claims of Daedalus Research that their system was accurate and reliable. Emerald Casino argued that the predictions were poor—half the people it classified as problem gamblers were not. Daedalus Research defended the system vigorously, arguing that their tests had shown it was 90% accurate and only 1-in-10 predictions were incorrect.

As is common practice in the ML community, Daedalus Research divided their data of 1000 people into training and test sets. Their algorithm was trained on the training set of 800 people, 400 problem and 400 non-problem gamblers.[160] It was then tested on the (up to then unseen) test set of 100 problem and 100 non-problem gamblers. It is common practice in the Machine Learning community for such an 80/20 split of training/test data. Daedalus Research reported 90% accuracy on this test set. That is, 180 of the 200 people in the test set were correctly identified as problem or non-problem gamblers, and just 20 of the 200 people in the test set were mis-classified.

The expert witness for the Emerald Casino pointed out the problem of considering just a simple summary statistic like accuracy and of the fact that in practice the problem is unbalanced — problem gamblers are typically in a minority compared to non-problem gamblers. Only around

[160] We suppose, in this hypothetical, that there is a reliable method to identify problem and non-problem gamblers that this machine learning system is trying to replicate. If the training data is not reliably labelled, then we are in the unfortunate position of "Garbage In, Garbage Out."

10% of the gambling population experience issues with their gambling. Thus, in a sample of 200 people, you might expect only about 20 problem gamblers, and not 100 as in the test set used by Daedalus Research. The expert witness for the Emerald Casino went on to note that a Machine Learning system that simply classified everyone as a non-problem gambler would achieve 90% accuracy but this is clearly not very useful.

Daedalus Research responded to these concerns by submitting a "confusion matrix" where the classification errors are broken out into false positives and false negatives (also called **type one** and **type two errors**), as well as true positives and true negatives. This data demonstrated that on the test set, the classifier was equally likely to give false positives as false negatives. That is, for the 20 people mis-classified, 10 people who were problem gamblers were classified as non-problem gamblers, and 10 people who were non-problem gamblers were classified as problem gamblers.

The system was thus 90% accurate at identifying non-problem gamblers correctly, and 90% accurate at identifying problem gamblers correctly. We can therefore estimate its accuracy on a representative sample of 200 people, 180 who are non-problem gamblers and 20 who are problem gamblers. 162 of these 180 non-problem gamblers (0.9 x 180) will be correctly classified as non-problem gamblers. And 18 of the 20 problem gamblers (0.9 x 20) will be correctly classified as problem gamblers. But 18 of the 180 (=180-162) non-problem gamblers will be incorrectly classified as problem gamblers. In total, 36 people (=18+18) people will be classified as problem gamblers, but 18 out of these 36 people classified as a problem gambler will not, in fact, be problem gamblers.[161] That is, as the Emerald Casino had claimed, half of the people classified by the classifier as a problem gambler were not problem gamblers.

A further concern raised by the expert witness from the Emerald Casino is "distributional shift." This is a change in the data distribution between an algorithm's training data, and the actual data encountered when deployed. [162] In this case, the training data was collected from the Emerald Casino in Hobart, Tasmania where, due to COVID restrictions, there are very few overseas visitors. However, when the system was applied to the Emerald Casino in Sydney, the data was very different due to the lifting of border restrictions and the presence of many more overseas visitors. Indeed, close analysis of the Hobart test set identified that. there, the classifier almost never identified overseas visitors as problem gamblers. As there were so few overseas visitors (in Hobart) in the training or test set, this had little impact on accuracy on the test set. By contrast, in the Sydney casino, half of all gamblers are from overseas, further degrading the accuracy and reliability of the classifier. It is not possible to quantify the amount

[161] For clarity: 36 is the total of people classified as problem gamblers; 18 are, in fact, problem gamblers and 18 are non-problem gamblers mis-classified.

[162] Put another way, a distributional shift is a change in the data distribution between an algorithm's training dataset, and a dataset it encounters when deployed (i.e., in the real world, a.k.a. the "wild"). Such shifts are common in practical applications of artificial intelligence.

by which performance degraded without data breaking down performance on overseas/non-overseas gamblers.

C.3. Conclusion / Sample Questions for Courts:

In considering the accuracy and reliability of an AI system, there are a range of issues that need to be considered. The following are sample questions courts may consider:

1. Was the dataset on which it was trained representative of the domain to which it was applied?
    - For instance, are the different classes (i.e., problem/non-problem gambler) balanced? How will this impact performance?
2. Are we trying to classify some rare event?
    - If so, we may need to consider performance very differently to events that are common.
3. Was the dataset "cleaned"?
    - Often, you will need to check for missing entries, erroneous data points and other anomalies in the data.
4. Did the data include all important features?
    - For instance, if gambling behavior of overseas visitors is very different to non-overseas visitors then this ought, probably, to be an input feature.
5. Was good practice used in training the system?
    - For example, was the data set separated into training and test set?
    - Was the data split between training and test set in a standard way (i.e. 80/20, 67/33, 50/50)?
6. Was performance analyzed carefully?
    - For example, were the different types of errors broken out? Perhaps the only errors are false positives and false positives are much more costly to fix than false negatives.
7. Was the model fixed or was it updated over time?
    - Once a model is deployed, you can expect distributional shift. It may be good practice to re-train the model at regular intervals to deal with such shift.