ADVISORY COMMITTEE
ON
EVIDENCE RULES

May 2, 2025

Reliability of AI-Generated Evidence

Proposed new Rule 707 aims to address the reliability of AI-generated evidence that is akin to expert testimony—and therefore comes with similar concerns about reliability, analytical error or incompleteness, inaccuracy, bias, and/or lack of interpretability. * * * Those concerns are heightened with respect to AI-generated content because it may be the result of complex processes that are difficult (if not impossible) to audit and certify. Examples of AI-generated evidence could include:

> • In a securities litigation, an AI system analyzes stock trading patterns over the last ten years to demonstrate the relative magnitude of the stock drop as a percentage of the Dow Jones Industrial Average, or to assess how likely it is that the drop in price was caused by a particular event.
>
> • An AI system analyzes keycard access records, iPhone GPS tracking, and Outlook calendar entries to demonstrate that an individual did not attend any of the senior management meetings over a period of time where alleged wrongdoing occurred.
>
> • In a copyright dispute, an AI system analyzes image data to determine whether two works are substantially similar.
>
> • An AI system assesses the complexity of an allegedly stolen software program in a trade secret dispute and renders an assessment of how long it would take to independently develop the code based on its complexity (and without the benefit of the allegedly misappropriated code).

Under the current rules, the methodologies that human expert witnesses employ and rely on are subject to Rule 702, which requires them to, among other things, establish that their testimony is based on sufficient facts or data; is the product of reliable principles and methods; and that those principles and methods are reliably applied to the facts of the case. See FRE Rule 702 (a)-(d). However, if machine or software output is presented on its own, without the

accompaniment of a human expert, Rule 702 isn't obviously applicable, see Reporter's Proposal at 51. This leaves courts and litigants to craft case-by-case frameworks for deciding when and whether AI-driven software systems can be allowed to make predictions or inferences that can be converted into trial testimony.

As a result, at its May 2, 2025 meeting, the Committee is expected to vote on proposed new Rule 707, Machine-Generated Evidence, drafted by the Committee's Reporter, Professor Daniel J. Capra of Fordham School of Law. (If approved, the Rule will be published for public comment.) The text of the proposed Rule provides:

> Where the output of a process or system would be subject to Rule 702 if testified to by a human witness, the court must find that the output satisfies the requirements of Rule 702 (a)-(d). This rule does not apply to the output of basic scientific instruments or routinely relied upon commercial software.

For instance, if a party uses AI to calculate a damages amount without proffering a damages expert, then they would need to prove that adequate data were used as the inputs for the AI program; that the AI program used reliable principles and methods; and that the resulting output is valid and reflects a reliable application of the principles and methods to the inputs, among other things. If adopted, Rule 707 analysis could require a determination of whether the training data is sufficiently representative to render an accurate output; whether the opponent and independent researchers have been provided sufficient access to the program to allow for adversarial scrutiny and sufficient peer review; and whether the process has been validated in sufficiently similar circumstances.

That the Committee is likely to approve this proposal underscores the federal judiciary's concerns about the reliability of certain AI-generated evidence that litigants have already sought to introduce in courtrooms. For example, U.S. District Judge Edgardo Ramos of the U.S. District for the Southern District of New York admonished a law firm for submitting ChatGPT-generated responses as evidence of reasonable attorney hourly rates because "ChatGPT has been shown to be an unreliable resource*." Z.H. v. New York City Dep't of Educ*., 2024 WL 3385690, at *5 (S.D.N.Y. Jul. 12, 2024). U.S. District Judge Paul Engelmayer similarly rejected AI-generated evidence because the proponent did "not identify the inputs on which ChatGPT relied" or substantiate that ChatGPT considered "very real and relevant" legal precedents. *J.G. v. New York City Dep't of Educ.,* 719 F. Supp. 3d 293, 308 (S.D.N.Y. 2024).

State courts also are beginning to grapple with the reliability of AI-generated evidence. For example:

In *Washington v. Puloka*, No. 21-1-04851-2 (Super. Ct. King Co. Wash. March 29, 2024), a trial judge excluded an expert's video where AI was used to increase resolution, sharpness, and definition because the expert "did not know what videos the AI-enhancement models are 'trained' on, did not know whether such models employ 'generative AI' in their algorithms, and agreed that such algorithms are opaque and proprietary." Id. at Par. 10.

In Matter of Weber as Tr. of Michael S. Weber Tr., 220 N.Y.S.3d 620 (N.Y. Sur. Ct. 2024), a New York state judge rejected a damages expert's financial calculations in part because he relied on Microsoft Copilot—a large language model generative AI chatbot—to perform calculations but could not describe the sources Copilot relied upon or how the AI tool arrived at its conclusion. In doing so, the judge reran the expert's inquiries on Copilot getting different results each time, and queried Copilot regarding its reliability, to which Copilot self-reported that it should be "check[ed] with experts for critical issues."

Reports indicate that a Florida state judge in Broward County recently donned a virtual reality headset provided by the defense to view a virtual scene of the crime from the perspective of the defendant who is charged with aggravated assault. The parties are likely to litigate the reliability of the technology before the judge decides if it can be used by a jury.

In both *Puloka* and *Weber,* the state courts emphasized that their respective jurisdictions follow the *Frye* standard, requiring scientific evidence to be generally accepted in its field, and found no evidence supporting the general acceptance of AI-generated evidence. These initial judicial reactions indicate that experts should be prepared to satisfy the jurisdiction-specific reliability standards for AI technologies they rely on when rendering their expert opinions.

Irrelevant material redacted

Irrelevant material redacted

X. Drafts of a New Rule 707

As stated above, this amendment treats the problem that arises where machine data would be considered expert testimony if coming from a person, but it is entered into evidence either directly or by someone who is not familiar with the machine's process and cannot verify its reliability. This problem arises most often today with attempts to "improve" visual or aural data by use of software that is not validated. What follows is: 1. the draft reviewed by the Committee at the last meeting, with suggested changes that are explained in the comments; and 2) A second draft that refers directly to "machine learning."

Rule 707. Machine-generated Evidence

Where the output of a process or system would be subject to Rule 702 if testified to by a human witness, the court must find that the output satisfies the requirements of Rule 702 (a)-(d). This rule does not apply to the output of basic scientific instruments ~~or routinely relied upon commercial software~~.

Comments:

1) The reference in the last sentence to routinely relied upon commercial software creates too broad an exclusion. For example, it could cover output from ChatGPT, if not now, then soon, because it will be "routinely relied upon." It can be argued that "basic scientific instruments," along with the Committee Note, will be sufficient guidance for courts in determining the scope of the rule. It is unlikely that any court is going to hold a *Daubert* hearing over a digital thermometer, regardless of what this rule says.

It could be further argued that the sentence should simply be *struck*, leaving the discussion of the breadth of the rule to the Committee Note. Again, one would not expect this rule to actually require a *Daubert* hearing for an electronic scale. But on the other hand, opponents may seek to exploit the lack of a limit in text.

The actual risks of overapplication of this rule will probably be raised in public comment. As such, for the public comment period, it is probably useful to have language in text for people to take a crack at. "Basic scientific instruments" is probably a good start for the comment period.

2) Another way to attempt to limit the rule is to put some qualifications on the term "process or system." If the goal of regulation is machines that learn things like humans, then perhaps the rule should be set forth as follows:

Where the output of a process or system **of machine learning** would be subject to Rule 702 if testified to by a human witness, the court must find that the output satisfies the requirements of Rule 702 (a)-(d).

This could be backed up by definition of machine learning in the Committee Note:

"Machine learning is an application of artificial intelligence that is characterized by providing systems the ability to automatically learn and improve on the basis of data or experience, without being explicitly programmed."

There would seem to be no risk of applying the rule to a digital thermometer if the scope of the rule is specifically limited to machine learning systems.

I ran this option by Professor Andrea Roth, who has graciously provided extremely valuable input to the Committee on this subject. Here is her answer:

"I think the term 'machine learning' describes a particular subset of algorithms that are 'trained' on data and then engage in either supervised or unsupervised 'learning' in terms of how to classify that data (what is a "dog" versus "cat," or what is this person's handwriting versus that person's handwriting, etc.). Deep neural networks and LLMs are a subset of machine learning that are particularly complex (involving "deep learning").

But an algorithm need not involve machine learning to be the sort of process or system that produces a machine-generated result and that would raise the issues underlying a proposed 707-like rule. For example, blood-alcohol software ... or Fitbit sleep tracking, gas chromatograph software, other forensic tools..."

So by using the term "machine learning" in the text the rule runs the risk of being underinclusive. But by covering all machines that would reach an expert-like conclusion, with a qualifying sentence at the end, you run the risk of being overinclusive. On balance, the risk of overinclusiveness may be the lesser risk; sensible courts are not going to conduct expert hearings on simple instruments. The risks of underinclusiveness are possibly greater because the line between a machine-learning process and other algorithmic calculations can be fuzzy, and is likely to become more fuzzy in the future. The current draft draws the line between *expert-like conclusions and non-expert-like conclusions.* And courts should be pretty good at assessing what would be an expert conclusion if coming from a human witness.

Just to show you what it would look like, there is a draft below (after the Committee Note) that is a machine-learning version of the rule.

Draft Committee Note

Expert testimony in modern trials increasingly relies on software- or other machine-based conveyances of information, from software-driven blood-alcohol concentration results to probabilistic genotyping ~~software~~. Machine-generated evidence can involve the use of a computer-based process or system to make predictions or draw inferences from existing data. When a machine draws inferences and makes predictions, there are concerns about the reliability of that process, akin to the reliability concerns about expert witnesses. Problems include using the process for purposes that were not intended (function creep); analytical error or incompleteness; inaccuracy or bias built into the underlying data or formulas; and

lack of interpretability of the machine's process. Where ~~an~~ a testifying expert relies on such a method, ~~the~~ that method – and the expert's reliance on it – will be scrutinized ~~pursuant to~~ under Rule 702. But if machine or software output is presented without the accompaniment of a human expert (for example through a witness who applied the program but knows little or nothing about its reliability), Rule 702 is not obviously applicable. Yet it cannot be that a proponent can evade the reliability requirements of Rule 702 by offering machine output directly, where the output would be subject to Rule 702 if rendered as an opinion by a human expert. Therefore, new Rule 707 provides that if machine output is offered ~~directly~~, without the accompaniment of an expert, its admissibility is subject to the requirements of Rule 702 (a)-(d).

The rule applies when machine-generated evidence is entered directly, but also when it is accompanied by lay testimony. For example, the technician who enters a question and prints out the answer might have no expertise on the validity of the output. Rule 707 would require the proponent to make the same kind of showing of reliability as would be required when an expert testifies on the basis of machine-generated information.

The rule is not intended to encourage parties to opt for machine-generated evidence over live expert witnesses. Indeed the point of the rule is to provide reliability-based protections when a party chooses to proffer machine evidence instead of a live expert.

It is anticipated that a Rule 707 analysis will usually involve the following, among other things:

• Considering whether the inputs into the process are sufficient for purposes of ensuring the validity of the resulting output. For example, the court should consider whether the training data for a machine learning process is sufficiently representative to render an accurate output for the population involved in the case at hand.

• Considering whether the process has been validated in circumstances sufficiently similar to the case at hand. For example, if the case at hand involves a DNA mixture of several contributors, likely related to each other, and a low quantity

of DNA, the software should be shown to be valid in those circumstances before being admitted.

The final sentence of the rule is intended to give trial courts sufficient latitude to avoid unnecessary litigation over ~~machine output that is regularly relied upon in commercial contexts outside litigation and that, as a result, is not likely to render output that is invalid for the purpose it is offered~~ the output from simple scientific instruments that are relied upon in everyday life. Examples might include the results of a mercury-based thermometer, an electronic scale, or a battery-operated digital thermometer. ~~or automated averaging of data in a spreadsheet, in the absence of evidence of untrustworthiness~~.

The Rule 702(b) requirement of sufficient facts and data, as applied to machine-generated evidence, should focus on the information entered into the process or system that leads to the output offered into evidence.

Comments:

1) There are a few refinements throughout, and an attempt to sharpen the paragraph that describes the "simple scientific instrument" exception. More examples of such instruments that are excluded from coverage can be added --- maybe as the result of public comment.

2) The paragraph on the risk that parties will not call experts but just admit machine data is addressed in a new paragraph, in response to the concerns of Judge Bates, expressed at the last meeting.

Draft Alternative --- Machine-Learning

Rule 707. Output of a Process of Machine-Learning

Where the output of a process or system of machine-learning would be subject to Rule 702 if testified to by a human witness, the court must find that the output satisfies the requirements of Rule 702 (a)-(d).

Draft Committee Note

Machine learning is an application of artificial intelligence that is characterized by providing systems the ability to automatically learn and improve on the basis of data or experience, without being explicitly programmed. Machine learning involves artificial intelligence systems that are used to perform complex tasks in a way that is similar to how humans solve problems. Machine-learning systems can make predictions or draw inferences from existing data supplied by humans. When a machine draws inferences and makes predictions, there are concerns about the reliability of that process, akin to the reliability concerns about expert witnesses. Problems include using the process for purposes that were not intended (function creep); analytical error or incompleteness; inaccuracy or bias built into the underlying data or formulas; and lack of interpretability of the machine's process. Where a testifying expert relies on the output of machine learning, that output – and the expert's reliance on it – will be scrutinized under Rule 702. But if machine learning output is presented without the accompaniment of a human expert (for example through a witness who applied the program but knows little or nothing about its reliability), Rule 702 is not obviously applicable. Yet it cannot be that a proponent can evade the reliability requirements of Rule 702 by offering machine learning output directly, where the output would be subject to Rule 702 if rendered as an opinion by a human expert. Therefore, new Rule 707 provides that if machine learning output is offered without the accompaniment of an expert, its admissibility is subject to the requirements of Rule 702 (a)-(d).

The rule applies when machine learning evidence is entered directly, but also when it is accompanied by lay testimony. For example, the technician who enters a question and prints out the answer might have no expertise on the validity of the output. Rule 707 would require the proponent to make the same kind of showing of reliability as would be required when an expert testifies on the basis of machine learning output.

The rule is not intended to encourage parties to opt for machine learning output evidence over live expert witnesses. Indeed the point of the rule is to provide reliability-based protections when a party chooses to proffer machine learning evidence instead of a live expert.

It is anticipated that a Rule 707 analysis will usually involve the following, among other things:

• Considering whether the inputs into the process are sufficient for purposes of ensuring the validity of the resulting output. For example, the court should consider whether the training data for a machine learning process is sufficiently representative to render an accurate output for the population involved in the case at hand.

• Considering whether the process has been validated in circumstances sufficiently similar to the case at hand. For example, if the case at hand involves a DNA mixture of several contributors, likely related to each other, and a low quantity of DNA, the software should be shown to be valid in those circumstances before being admitted.

The Rule 702(b) requirement of sufficient facts and data, as applied to machine learning evidence, should focus on the information entered into the process or system that leads to the output offered into evidence.